REVISIÓN SISTEMÁTICA PARA LAS TÉCNICAS DE MINERÍA WEB DE CONTENIDO

Jehison David Cifuentes Cortés
Jorge Mauricio Martínez Naizaque

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
TRABAJO DE GRADO
BOGOTÁ, D.C.
2018

REVISIÓN SISTEMÁTICA PARA LAS TÉCNICAS DE MINERÍA WEB DE CONTENIDO

Jehison David Cifuentes Cortés
Jorge Mauricio Martínez Naizaque

Trabajo de grado presentado como requisito parcial para optar al título de:
Ingeniero de Sistemas

Director
Ing. Nelson Augusto Forero Páez

UNIVERSIDAD CATÓLICA DE COLOMBIA
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
TRABAJO DE GRADO
BOGOTÁ, D.C.
2018

NOTA DE ACEPTACIÓN

Aprobado por el comité de grado en cumplimiento de los requisitos exigidos por la Facultad de Ingeniería y la Universidad Católica de Colombia para optar por el título de Ingeniero de Sistemas.

_____
Jurado

_____
Ing. Nelson Augusto Forero Páez
Director

_____
Revisor metodológico.

Bogotá, D.C., mayo 21 de 2018

AGRADECIMIENTOS

En primer lugar, queremos agradecer a Dios por permitirnos desarrollar este trabajo y ser nuestra fortaleza en los momentos de dificultades, así como brindarnos la grandiosa oportunidad de ser parte de nuestra Universidad Católica de Colombia.

A nuestro director de trabajo de grado, *Ingeniero Nelson Augusto Forero Páez*, por su esfuerzo y dedicación, quien con sus conocimientos, experiencia, motivación y paciencia ha orientado nuestro trabajo hasta que conseguimos llevarlo a feliz término, con éxito.

También nuestra gratitud es para todos los docentes que durante el transcurso de la carrera profesional nos aportaron conocimientos, experiencia y consejos, con los que contribuyeron de forma invaluable a nuestra formación como personas y como profesionales, en especial, los *Ingenieros Jorge Carrillo, Holman Bolívar, JJ, Diego Rincón, Diego Velandia y Raúl Menéndez.*

Y, por último, a nuestra alma mater, Universidad Católica de Colombia, institución que nos brindó la oportunidad, a través de su programa de Ingeniería de Sistemas, de realizar nuestros estudios y de la cual siempre recibimos apoyo.

Jehison David
Jorge Mauricio

CONTENIDO

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág.

GLOSARIO

BIG DATA: Big data, macro datos o datos masivos es un concepto que hace referencia al almacenamiento de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos. El fenómeno del Big data también se denomina a veces datos a gran escala. En los textos científicos en español con frecuencia se usa directamente el término en inglés Big data, tal como aparece en el ensayo seminal de Viktor Schönberger big data: La revolución de los datos masivos[1].

CACHES BUSTING: Técnica para garantizar que los navegadores o servidores Proxy siempre obtengan una copia nueva de la petición realizada al sitio Web, evitando obtener copias a partir de otras caches[2].

COOKIE: Es un archivo que se almacena en el disco duro del visitante de una página Web a través de su navegador, a petición del servidor de la página. Esta información es recuperada por el servidor en posteriores visitas. Las inventó Lou Montulli, un antiguo empleado de Netscape Communications[3].

CLUSTER: (a veces castellanizado como clúster) es un término inglés encontrado en varios tecnicismos. La traducción literal al castellano es "racimo", conjunto, "grupo" o "cúmulo":
Se aplica a los conjuntos o conglomerados de computadoras construidos mediante la utilización de hardwares comunes y que se comportan como si fuesen una única computadora[4].

CLUSTERING: También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares. Cuando se representan la información obtenida a través de clusters se pierden algunos detalles de los datos, pero a la vez se simplifica dicha información[5].

DIAGRAMA DE CASO DE USO: Es un tipo de clasificador representando una unidad funcional coherente, un subsistema o una clase manifestada por secuencias de mensajes[6].

---

[1] VILLATE, Jaime. Glosario de informática Inglés-Español. 2000. Recuperado de: http://quark.fe.up.pt/orca/pub-es/glosario.html
[2] Ibíd.
[3] Ibíd.
[4] Ibíd.
[5] Ibíd.
[6] Ibíd.

DIAGRAMA DE CLASES: Es el diagrama principal para el análisis y diseño. Un diagrama de clases presenta las clases del sistema con sus relaciones estructurales y de herencia. La definición de clase incluye definiciones para atributos y operaciones. El modelo de casos de uso aporta información para establecer las clases, objetos, atributos y operaciones[7].

FIREWALL Un firewall es un elemento de hardware o software utilizado en las redes para prevenir algunos tipos de comunicaciones prohibidas por las políticas de red, las cuales se fundamentan en las necesidades del usuario[8].

HITS: Un hit es un acceso, una petición al servidor de un fichero; por ejemplo, si en una página, además del archivo php, usó un archivo externo javascript, otro css, y, además, la cabecera y 5 imágenes pequeñas, resulta que se tienen 9 hits, nueve peticiones de ficheros[9].

JDBC: Acrónimo de Java Database Conectivity, es un API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java independientemente del sistema de operación donde se ejecute o la base de datos a la cual se accede utilizando el lenguaje SQL del modelo de base de datos[10].

KDD: El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos)[11].

LATENCIA: Tiempo en que tardan en comunicarse dos puntos remotos[12].

LOG: En informática, se usa el término log, historial de log o registro a la grabación secuencial en un archivo o en una base de datos de todos los acontecimientos (eventos o acciones) que afectan a un proceso particular (aplicación, actividad de una red informática, etc.). De esta forma constituye una evidencia del comportamiento del sistema[13].

MÁQUINAS DE APRENDIZAJE: Es un área de la inteligencia artificial concerniente al desarrollo de técnicas que permiten a las computadoras “aprender”. Es un método

[7] Ibíd.
[8] Ibíd.
[9] Ibíd.
[10] Ibíd.
[11] Ibíd.
[12] Ibíd.
[13] Ibíd.

para crear programas de computadora orientados al análisis de conjuntos de datos[14].

MODELO ENTIDAD-RELACIÓN: Es el modelo conceptual más utilizado para el diseño conceptual de bases de datos. Fue introducido por Peter Chen en 1976. El modelo entidad-relación está formado por un conjunto de conceptos que permiten describir la realidad mediante un conjunto de representaciones gráficas y lingüísticas. Originalmente, el modelo entidad-relación sólo incluía los conceptos de entidad, relación y atributo. Más tarde, se añadieron otros conceptos, como los atributos compuestos y las jerarquías de generalización, en lo que se ha denominado modelo entidad-relación extendido[15].

PAGERANK: Es una marca registrada y patentada por Google el 9 de enero de 1999 que ampara una familia de algoritmos utilizados para asignar de forma numérica la relevancia de los documentos (o páginas web) indexados por un motor de búsqueda. Sus propiedades son muy discutidas por los expertos en optimización de motores de búsqueda[16].

PRECARGA: En término de computación se refiere a la práctica que consiste en cargar a memoria datos necesarios para desempeñar ciertas tareas computacionales antes de que estas inicien[17].

RECONOCIMIENTO DE PATRONES: Es un área incluida dentro de las máquinas de aprendizaje que se enfoca en clasificar datos basándose en conocimiento previo o información estadística previamente extraída a partir de los patrones[18].

ROBOT DE INTERNET: Los robots, en Internet, son también conocidos como arañas, y se trata de programas que navegan, por su cuenta, y por medio de programación en el tiempo, con el objeto de visitar sitios y obtener información de éstos[19].

SERVIDOR PROXY: El término proxy hace referencia a un programa o dispositivos que realiza una acción en representación de otro. La finalidad más habitual de esa representación es la de permitir el acceso a Internet a todos los equipos de una organización cuando sólo se puede disponer de un único equipo conectado, esto es, una única dirección IP[20].

---

[14] Ibíd.
[15] Ibíd.
[16] Ibíd.
[17] Ibíd.
[18] Ibíd.
[19] Ibíd.
[20] Ibíd.

SERVIDOR WEB: Un servidor Web es un programa que implementa el protocolo http (hypertext transfer protocol). Este protocolo está diseñado para transferir lo que llamamos hipertextos, páginas Web o páginas HTML (hypertext markup language)[21].

URI: Uniform Resource Identifier, es decir, identificador uniforme de recursos. Texto corto que identifica unívocamente cualquier recurso (servicio, página, documento, etc.) accesible en una red[22].

URL: Uniform Resource Locutor, es decir localizador uniforme de recurso. Es la cadena de caracteres con la cual se asigna dirección única a cada uno de los recursos de información disponible en Internet[23].

VISITAS: Cuando un internauta entra en una página es una visita. Todo el tiempo que navegue por dicho sitio Web contará como una visita, sólo una; la primera petición que realiza ese cliente remoto, es lo que cuenta como visita, el tiempo que pase en la Web, descargando algo, leyendo contenidos, todo eso formará parte de la misma visita[24].

WEBCRAWLER: Es un metabuscador que combina la búsquedas tope de Google, Yahoo!, Bing (antes MSN Search), Ask.com, About.com, MIVA, LookSmart y otros motores de búsqueda populares. WebCrawler también proporciona a los usuarios la opción de búsqueda de imágenes, audio, vídeo, noticias, páginas amarillas y páginas blancas. WebCrawler es una marca registrada de InfoSpace. Inc[25].

WORLD WIDE WEB: Es básicamente un medio de comunicación de texto, gráficos y otros objetos multimedia a través de Internet, es decir, la web es un sistema de hipertexto que utiliza Internet como su mecanismo de transporte o desde otro punto de vista, una forma gráfica de explorar Internet[26].

[21] Ibíd.
[22] Ibíd.
[23] Ibíd.
[24] Ibíd.
[25] Ibíd.
[26] Ibíd.

RESUMEN

El estudio tuvo como objetivo determinar las principales técnicas empleadas de minería web que permiten realizar minería de contenido, con el fin de facilitar la búsqueda de información en bases documentales. Para ello se llevó a cabo una revisión sistemática de la información documentada en medios arbitrados en el período 2014 – 2018, empleando las bases documentales Redalyc, Scielo, Scopus, IEEEXplore, Google Scholar y Web of Science, con términos de búsqueda como “Minería web”, “Contenido web”, “Técnicas de minería web”, “Web mining", "Web content", "Web mining techniques”, y aplicando filtros de año de publicación y texto completo, a partir de lo cual se identificaron inicialmente 49231 documentos que fueron revisados a la luz de los criterios de inclusión. De ahí se seleccionaron y analizaron los artículos a texto completo publicados entre el 2014 y el 2018 con información relevante sobre las principales técnicas empleadas de minería web de contenido que facilitan la búsqueda de información en bases documentales. Se obtuvo como resultado de la revisión que las técnicas de minería web de contenido se categorizan en no estructuradas, estructuradas, semi estructuradas y multimedia, las cuales se complementan con las técnicas de minería de datos en las cuales las más representativas fueron las predictivas y las descriptivas.

Palabras clave: Análisis de contenido, análisis documental, minería, técnicas.

ABSTRACT

The objective of this study was to determine the main techniques employed in web mining that allow content mining, in order to facilitate the search for information on documentary bases. For this, a systematic review of the information documented in arbitrated media in the period 2014 - 2018 was carried out, using the documentary databases Redalyc, Scielo, Scopus, IEEEXplore, Google Scholar and Web of Science, with search terms such as “Minería web”, “Contenido web”, “Técnicas de minería web”,"Web mining", "Web content", "Web mining techniques", and applying year-of-publication and full-text filters, from which 49231 documents were initially identified that were reviewed in light of the inclusion criteria. From there, the full-text articles published between 2014 and 2018 were selected and analyzed with relevant information on the main techniques employed in content web mining that facilitate the search for information on documentary bases.
It was obtained as a result of the review that web content mining techniques are categorized into unstructured, semi-structured and multimedia structures, which are complemented by data mining techniques in which the most representations were predictions and descriptive ones.

Keywords: Content analysis, documentary analysis, mining, techniques.

INTRODUCCIÓN

Los altos volúmenes de información que se producen y publican a diario en la web hacen que se aumente su disponibilidad, pero al mismo tiempo dificultan el acceso a ella, en la medida que es requerida. Por esa razón se vuelve muy importante identificar técnicas de minería web que permitan realizar búsquedas de contenido en la web, con el fin de facilitar la búsqueda de información en bases documentales.

Con ello como propósito se planteó un estudio de revisión sistemática en el que se contempla el diseño metodológico propio de esta metodología efectuando el análisis de documentos publicados entre los años 2014 y 2018 procedentes de seis bases documentales, a partir de los cuales se logró recopilar información relevante a las principales técnicas empleadas en minería web de contenidos.

Así, se estructuró un documento que presenta en su primer capítulo las generalidades del estudio que incluyen los antecedentes, el planteamiento del problema, los objetivos trazados, la justificación, la delimitación, el marco referencial, la metodología empleada y el diseño metodológico; en el segundo capítulo la definición de los criterios de búsqueda de artículos sobre minería de contenido en la web; en el tercer capítulo la identificación de las principales técnicas que han sido empleadas para realizar minería de contenido en la web; en el cuarto capítulo la caracterización de dichas técnicas de minería web identificadas; y en el quinto capítulo, las conclusiones y recomendaciones resultantes del estudio.

1. GENERALIDADES

1.1 ANTECEDENTES

Los antecedentes de la minería web se remontan a la década de los años 90's y primeros años del siglo XXI, épocas en las que se presentaron los inicios del manejo virtual de la información, razón por la cual los referentes que se emplean a continuación, inicialmente datan de esos tiempos.

Dado el crecimiento exponencial de la información disponible en la web, la World Wide Web, se ha convertido en una plataforma poderosa para almacenar, dispersar y recuperar la información, proceso que se conoce como minería Web.
La minería web tiene su origen como concepto, en la minería de datos. La minería web se concibe como una minería de datos, pero aplicada en la www. Los orígenes de la minería de datos se remontan hasta finales de los años ochenta cuando el término comenzó a ser usado sobre todo en el campo de la investigación.[27]

En los años noventa se conocía comúnmente como un subproceso dentro de un proceso más grande que era el descubrimiento de conocimiento en bases de datos o Knowledge Discovery in Databases.[28]

A partir del año 2000, con el auge de las redes sociales y la expansión de la web 2.0, el análisis de sentimientos se ha convertido en un tópico de interés como resultado del procesamiento del lenguaje natural. Para realizar tareas como la mencionada se requiere de la extracción de información a partir de la WEB. Con base en la información obtenida, algoritmos como el Page Rank de Google pueden recuperar páginas web y crear una lista considerando el valor de relevancia de las páginas recuperadas con respecto a una consulta suministrada.

Por lo tanto, la minería web incluye el descubrimiento y análisis de información relevante, a partir del uso de técnicas basadas en minería de datos, orientados al descubrimiento y extracción automática de información de documentos y servicios ofrecidos por la web.[29]

---

[27] FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (ed.) *Advances in knowledge and data mining*. Cambridge (Massachussets): AAAI/MIT Press, 1996.
[28] MOLINA, L.C. *Torturando los Datos Hasta que Confiesen*. Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña. Barcelona, España, 2000
[29] SCOTTO, M.; SILLITTI, A.; SUCCI, G.; VERNAZZA, T. *"Managing Web-Based Information",* International Conference on Enterprise Information Systems (ICEIS 2004), Porto, Portugal, April 2004. Page 1-3

1.2 PLANTEAMIENTO DEL PROBLEMA

1.2.1 Descripción del problema. La web almacena un gran volumen de información, esto dificulta al usuario acceder de manera ágil a aquella que le sea relevante o importante, dado que no hay un patrón que permita un fácil acceso. Además, en algunas ocasiones la información y los datos a pesar de estar aparentemente al alcance de la mano, se encuentran en diferentes tipos de formatos que no siempre son de fácil manipulación, incrementando la dificultad mencionada de acceso.

Para ello se requiere de técnicas que permitan y faciliten el acceso a la información de los contenidos de los sitios web, las cuales se han ido desarrollando en el transcurso del tiempo, pero con poca difusión, lo que las hace desconocidas para la mayoría de personas. Estas técnicas corresponden a lo que se ha denominado como minería web.

La minería web puede dividirse en tres tipos, uso de la web, contenido de la web y estructura de la web. La minería web de uso se focaliza en los recursos que se consumen en las páginas web, trata de determinar qué secciones o categorías presentan mayor relevancia para los usuarios, patrones de acceso, y navegación de usuarios en sitios y páginas.[30]

La minería web de estructura busca extraer información de los hiperenlaces, debido a que cada página o sitio puede referenciar a través de hipervínculos otras páginas, es posible extraer información útil al explorar el grafo de enlaces de la web.[31]

Por último, la minería web de contenido extrae información útil desde los contenidos de las páginas web disponibles. Las páginas web contienen objetos con contenido en diversos formatos como texto, audio, imágenes o video.[32] Dado que la minería web de contenido en la mayoría de los estudios consiste en la exploración de patrones de interés en texto, es frecuente que esta tarea sea asociada a minería de texto.[33] En general la minería web de contenido explora el texto. Esto debido a que la exploración del contenido en otros formatos como audio o video, es en general una tarea difícil.[34]

---

[30] MENDOZA, M. Minería de datos en la web. Capítulo 19, 613-648, en CACHEDA, F.; FERNÁNDEZ, J.; y HUETE, J. Recuperación de información: Un enfoque práctico y multidisciplinar, Editorial Ra-Ma, 2011.
[31] Ibíd.
[32] Ibíd.
[33] Ibíd.
[34] Ibíd.

1.2.2 Formulación del problema. Conforme a esto nace la minería web como un proceso global para descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la web.[35]

Con base en lo anterior se planteó la siguiente pregunta de investigación:
¿Cuáles son las principales técnicas empleadas que permiten realizar minería de contenido en la web, para facilitar la búsqueda de información en bases documentales?

1.3 OBJETIVOS

1.3.1 Objetivo general. Determinar las principales técnicas de minería web que permiten realizar minería de contenido en la web, con el fin de facilitar la búsqueda de información en bases documentales.

1.3.2 Objetivos específicos. El trabajo se llevó a cabo mediante el desarrollo de los siguientes objetivos específicos:

- Definir los criterios de búsqueda de artículos sobre minería de contenido en la web a ser empleados en las bases documentales, mediante la elaboración de una estrategia de búsqueda.

- Identificar las técnicas que han sido empleadas para realizar minería de contenido en la web, mediante el análisis de los artículos seleccionados.

- Caracterizar las principales técnicas para realizar minería de contenido en la web, a partir del análisis de los artículos seleccionados.

1.4 JUSTIFICACIÓN

Teniendo en cuenta el tamaño de la web, se ha hecho cada vez más necesario para los usuarios y para las empresas utilizar herramientas automatizadas para encontrar y recuperar los recursos de información deseados.

El poder de la información en la era digital se ha hecho parte fundamental para el mundo de los negocios. La mayoría de las empresas compiten en cuanto al uso, tratamiento y aprovechamiento que les dan a los datos, y consideran el poder del conocimiento como una ventaja competitiva, reforzando sus estrategias de negocio en algunos aspectos como, innovación de sus procesos y operaciones, orientar esfuerzos hacia los clientes con un mayor valor, mejorar las experiencias para generar lealtad de los clientes hacia sus productos o servicios, generar productos o servicios más competitivos en el mercado.

[35] SRIVASTAVA, J.; COOLEY, R.; DESHPANDE, M. and TAN, P.N. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, I(2):12-23, 2000.

Por ejemplo, tener dominio sobre la información puede generar cierto tipo de ventajas como en el mercado de las divisas, a un inversor a largo plazo, la información puede servirle como herramienta para determinar hacia donde se dirige el mercado, donde se va a posicionar en los meses o años próximos.

En fin, las principales técnicas que permiten realizar minería de contenido en la web, para facilitar la búsqueda de información en bases documentales, es totalmente relevante no solo por la necesidad de contar con la información requerida en el momento oportuno sino porque ello implica, además, el mejor aprovechamiento del recurso tiempo y del recurso financiero en las organizaciones.

1.5 DELIMITACIÓN

1.5.1 Espacio. El presente proyecto se realizó en Colombia donde se llevó a cabo tanto el trabajo de investigación, propuestas, resultados obtenidos y el trabajo de revisión sistemática.

1.5.2 Tiempo. El tiempo de realización de este proyecto con base a la planificación realizada por la Universidad Católica de Colombia inició el 10 de febrero de 2018 y terminó el 30 de abril de 2018 con la entrega del informe final.

1.5.3 Contenido. El contenido del proyecto consistió en realizar una investigación sobre las principales técnicas que permiten realizar minería de contenido en la web.

1.5.4 Alcance. El alcance del estudio es el desarrollo de un documento con la descripción de las principales técnicas de minería web que permiten realizar minería de contenido en la web, con el fin de facilitar la búsqueda de información en bases documentales.

1.6 MARCO TEÓRICO

Una revisión sistemática de la literatura permite identificar, evaluar, interpretar y sintetizar todas las investigaciones existentes y relevantes en un tema de interés particular. Este tipo de revisiones se ejecutan de forma rigurosa e imparcial para que tengan un valor científico. La principal motivación para emprender una revisión sistemática es incrementar la posibilidad de detectar más resultados reales en el tema de interés que los que pueden ser detectados con revisiones de menor dimensión[36].

[36] Íbid

1.6.1 Internet. La internet se concibe como una red integrada por miles de redes y computadoras interconectadas en todo el mundo mediante cables y señales de telecomunicaciones, que utilizan una tecnología común para la transferencia de datos, en donde el protocolo que emplea se denomina TCP/IP (Protocolo de control de transmisión / Protocolo de Internet)[37].

1.6.2 La dirección IP. La dirección IP o dirección de Internet es “una dirección de origen o destino de 4 octetos (32 bits) formada por un campo de Red y un campo de Dirección Local". Ello denota que la finalidad del establecimiento de estas direcciones es el reconocer máquinas interconectadas a través del protocolo IP, y no necesariamente a las personas que están operándolas. Sin embargo, con la masificación de Internet y su desarrollo como sistemas de intercambio de la información, se ha cuestionado la posibilidad de que estas direcciones sean consideradas un dato personal[38].

1.6.3 World Wide Web. La web, como se le suele conocer, se considera el mayor portento tecnológico que el hombre haya desarrollado. Su impacto en la sociedad ha sido tal que se le ha comparado con la invención de la rueda o el descubrimiento del fuego[39].

Desde los orígenes de la Web, la creación de un sitio no ha sido un proceso fácil. Muchas veces se requiere de un equipo multidisciplinario de profesionales enfocados a una sola misión: asegurar que el contenido y la estructura del sitio le sean atractivos al usuario. Lo anterior se ha abordado con relativo éxito en el ámbito de la *personalización de la Web*, concepto que es la clave del éxito para obtener una adecuada participación en el mercado electrónico, mantener la vigencia del sitio y, sobre todo, lograr la fidelización del cliente digital[40].

En esencia, los algoritmos, técnicas y métodos que comprende el web mining, son utilizados en el procesamiento masivo de datos, lo cual requiere una automatización parcial o total de todas las operaciones a fin de obtener los resultados en cuestión de horas o días. En consecuencia, el análisis de la web data, utilizando técnicas de web mining, cuenta con todos los requisitos necesarios para ser estudiado a partir de la regulación nacional e internacional que hasta el momento se ha desarrollado[41].

---

[37] ZAMORA, M.A. Internet. Universidad Autónoma del Estado de Hidalgo. 2014. P.3

[38] Ibíd. p.53

[39] BERNERS, T.; CAILLIAU, R.; LUOTONEN, A.; NIELSEN, H. F. & SECRET. A. The world wide web. Communications of ACM, 37(8):76-82, 1994.

[40] VELÁSQUEZ, J.D. & DONOSO, L. Aplicación de Técnicas de Web Mining sobre los Datos Originados por Usuarios de Páginas Web. Visión Crítica desde las Garantías Fundamentales, especialmente la Libertad, la Privacidad y el Honor de las Personas. Revista Ingeniería de Sistemas, XIV: 47-68, Junio 2010.

[41] CARRASCO-JIMÉNEZ, P. Análisis Masivo de Datos y Contraterrorismo. Tirant lo Blanch, Valencia, España, 2009.

1.6.4 Internet y web. La Web e Internet son conceptos distintos pero que a menudo se confunden. Internet representa a la red de redes que permite la interconexión de dispositivos que se encuentran a nivel local, con sus similares en una región diferente, a través del envío y recepción de los datos que viajan en paquetes. La Web es el conjunto de páginas y objetos relacionados que se vinculan entre sí a través de hipervínculos. A un conjunto de páginas web se le denomina sitio web y es administrado por una aplicación conocida como servidor web, la cual utiliza a Internet como lugar físico para transferir las páginas web y otros objetos asociados[42].

1.6.5 Datos originados en la web. Cada una de las páginas posee un contenido representado a través de objetos como texto, imágenes, sonidos, películas o vínculos a otros sitios web. Su funcionamiento se describe a continuación: El servidor web o web server es una aplicación que está en ejecución continúa atendiendo requerimientos de objetos web, es decir, el conjunto de archivos que conforman el web site y los envía a la aplicación que hace la solicitud, generalmente un web browser. En general estos archivos son imágenes, sonidos, películas y páginas web que conforman la información visible del sitio. Las páginas están escritas en Hyper Text Markup Language (HTML), que en síntesis es un conjunto de instrucciones, también conocidas como tags, acerca de cómo desplegar objetos en el browser o dirigirse a otra página web (hyperlinks). Estas instrucciones son interpretadas por el browser, el cual muestra los objetos en la pantalla del usuario[43].

Cada uno de los tags presentes en una página, son interpretados por el browser. Algunos de estos hacen referencia a otros objetos en el web site, lo que genera una nueva petición en el browser y la posterior respuesta del server. En consecuencia, cuando el usuario digita la página que desea ver, el browser, por interpretación secuencial de cada uno de los tags, se encarga de hacer los requerimientos necesarios que permiten bajar el contenido de la página al usuario[44].

1.6.6 Minería de datos en la Web. Web mining es el concepto que agrupa a todas las técnicas, métodos y algoritmos utilizados para extraer información y conocimiento desde los datos originados en la Web (web data). Parte de estas técnicas apuntan a analizar el comportamiento de los usuarios, con miras a mejorar continuamente la estructura y contenido de los sitios que son visitados.

Para ayudar al usuario a que se sienta lo mejor atendido posible por el sitio web, en han desarrollado una serie de metodologías para el procesamiento de datos, cuya

[42] VELÁSQUEZ, J.D. & DONOSO, L. Op.cit. p.49
[43] COOLEY, R. MOBASHER, B. AND SRIVASTAVA. J. Data preparation for mining world wide web browsing patterns. Journal of Knowlegde and Information Systems, 1(1):5-32, 1999.
[44] VELÁSQUEZ, J.D. & DONOSO, L. Op.cit. p.50

operación es al menos cuestionable, desde el punto de vista de la privacidad de los usuarios de un sitio web determinado[45,46].

1.6.7 Limpieza y procesamiento de los web data. Los datos originados en la Web o web data, corresponden esencialmente a tres fuentes[47]:
1. Contenido: Son los objetos que aparecen dentro de una página web, por ejemplo, las imágenes, los textos libres, sonidos, etc.
2. Estructura: Se refiere a la estructura de hipervínculos presentes en una página.
3. Uso: Son los registros de web logs, que contienen toda la interacción entre los usuarios y el sitio web.

Los web data deben ser pre procesados antes de entrar en un proceso de web mining, es decir, son transformados en vectores de características que almacenan la información intrínseca que hay dentro de ellos[48,49].

Aunque todos los web data son importantes, especial atención reciben los web logs, ya que ahí se encuentra almacenada la interacción usuario sitio web, sus preferencias de contenido y en síntesis su comportamiento en el sitio.

1.6.8 Criterios para seleccionar contenidos web. Los siguientes son los criterios a tener en cuenta cuando se está efectuando la selección de contenidos web[50]:

URL: ¿Qué sugiere la dirección de la página? ¿Cuál es el dominio principal? ¿el último código a la derecha de la primera parte de la URL? ¿O el subdominio? ¿Algún término significativo? ¿Deduce si la raíz es de una web educativa, oficial o comercial? Lo educativo u oficial inspira confianza: ¿dónde lo clasifica? ¿Otros indicios?[51]
• .com: sitios de empresas, con intereses comerciales.
• .net: sitios de empresas tecnológicas o de comunicaciones.
• .edu: sitios educativos, de investigación, etc. en USA y en muchos países.
• .ac.*: sitios educativos, de investigación, etc. en países anglosajones.
• .edu.*: sitios educativos, de investigación, etc. en países iberoamericanos, etc.
• .org: sitios de entidades sin fines lucrativos, pueden tener intereses ideológicos.

---

[45] TAVANI, H.T. Informational privacy, data mining, and the internet. Ethics and Information Technology, 1:137-145, 1999.
[46] VEDDER, A. Privacy and con dentiality. medical data, new information technologies, and the need for normative principles other than privacy rules. Law and Medicine, 3:441{459, 2000.
[47] COOLEY, R. MOBASHER, B. AND SRIVASTAVA. J. Op. cit.
[48] EIRINAKI, M.; and VAZIRGANNIS, M. Web mining for web personalization. ACM Transactions on Internet Technology, 3(1):1-27, February 2003.
[49] VELÁSQUEZ, J.D. and PALADE, V. A knowledge base for the maintenance of knowledge extracted from web data. Knowledge-Based Systems, 20(3):238-248, 2007.
[50] MARTÍNEZ, L.J. Cómo buscar y usar información científica: Guía para estudiantes universitarios. Santander, España: Universidad de Cantabria, 2013.
[51] Ibíd, p.18

• .es, .uk, .fr, .pt, .de, .mx, .ar, .co, .cl, .pe: son dominios geográficos.
• ~: como parte de la URL es indicio de que se trata de una página personal.

Sitio web: ¿Cuál es el sitio web donde se aloja el contenido que se examina? ¿Qué confianza aporta? Pulse el enlace Inicio, Home, etc. para verlo, o borre con el cursor en la barra de dirección del navegador hasta la página principal. ¿A quién pertenece la web? ¿Qué institución está detrás? ¿Qué propósitos le mueven? Busque en los enlaces Presentación, Acerca de, About us, Objetivos, etc. ¿Los webmaster filtran y controlan los contenidos de la web o es un sitio donde usuarios externos autopublican contenidos sin más?[52]

Autoría: ¿Figuran los responsables directos de la creación del contenido? Esto sería buen síntoma, cuanto más anónima la información, peor. A veces, no obstante, la autoría puede no ser personal, sino de grupos, colectivos o instituciones. ¿Son expertos en la materia? ¿Constan sus datos, sus credenciales, curriculum, forma de contacto, etc.? Tenga en cuenta que busca información científica, de expertos, no de aficionados, ni de personas que sepan menos o igual que usted[53].

Vigencia: ¿La información está datada, incluye fecha? Esto es otro buen síntoma, en sí mismo. Por la fecha, y por el tema o rama de conocimiento, ¿puede considerarla vigente u obsoleta? ¿Corre riesgos? Es muy importante. Puede haber otros indicios en el texto: fechas citadas, noticias, datos, legislación, referencias bibliográficas con año, etc. Aunque es un indicio indirecto, tampoco hay que confundir la fecha de actualización de la web con la vigencia o actualización del contenido[54].

Finalidad: ¿Para qué y para quién está pensada la página web en cuestión? ¿A quién se dirige? ¿Con qué propósito? Y, por ende, ¿qué nivel intelectual alcanza? ¿Se adapta a la exigencia de sus necesidades y requerimientos? A título de ejemplo, pregúntese[55]:
• ¿Son resultados de la investigación, para otros investigadores?
• ¿Es información para profesionales, especialistas, expertos?
• ¿Es un material educativo, formativo? ¿De qué nivel de enseñanza?
• ¿Es información comercial de una empresa para potenciales clientes?
• ¿Son opiniones de/para aficionados, interesados, afectados, partidarios…?
• ¿Es un contenido sólo para generar tráfico hacia los anuncios?
• ¿Es divulgación científica? ¿De qué nivel y pretensiones parece?
• ¿Es información de la administración para el ciudadano?

[52] Ibíd, p.18
[53] Ibíd, p.18
[54] Ibíd, p.18
[55] Ibíd, p.18

Rigor: ¿El texto parece redactado de forma apropiada? ¿Usa un lenguaje científico preciso? ¿Expone correctamente la información? ¿Justifica sus afirmaciones mediante referencias bibliográficas? ¿Cita otros estudios o informes, aporta documentación? ¿Incluye datos: experimentos o cálculos propios, cifras tomadas de fuentes ajenas, ¿etc.? ¿Ofrece enlaces vivos a otros sitios web? ¿Transmite seguridad?[56]

Consistencia: La información, ¿incluye contradicciones internas?, ¿Tiene afirmaciones sospechosas, contradice algo que sepa?, ¿Incorpora enunciados que en otras fuentes figuran de otra forma?[57]

Objetividad: Es importante vigilar si hay sesgos ideológicos, o de otro tipo. ¿Pretenden vender algo: ¿una idea, un producto? ¿La información es tendenciosa? ¿Cuál es el balance entre persuasión, opinión e información? ¿Hay intereses, ocultos o visibles? ¿La publicidad afecta al contenido?[58]

Diseño: El diseño puede decir mucho. ¿Está bien cuidado o es desaliñado? ¿Antiguo o moderno? Sobre todo: ¿la información está bien organizada y estructurada? ¿Qué domina: textos o imágenes? ¿Es llamativo, para captar la atención, o austero? La información más seria y fiable tiende a estar bien presentada, pero suele ser sobria. ¿Hay publicidad? ¿Mucha, invasiva? El exceso y preeminencia de la publicidad revela poca consideración del valor de la información por parte del webmáster: es mala señal[59].

Relevancia: Es vital tener esto en cuenta al valorar un resultado de una búsqueda. ¿La información es pertinente para lo que se busca? ¿Responde a sus preguntas? ¿Tiene que ver con su necesidad?[60]

Suficiencia: La página o contenido, ¿qué cantidad de información le aporta en relación con su problema? ¿Es suficiente para lo que busca? Una sola fuente nunca suele serlo, pero ¿abarca todos los aspectos del tema? ¿Con qué grado de detalle, de profundidad?[61]

Conclusión: Valorando en resumen todos los aspectos anteriores, ¿qué opinión le suscita la página que analiza? ¿Merece confianza? ¿Es apropiada para el usuario? ¿Es adecuada para su necesidad? ¿Tiene suficiente fiabilidad, credibilidad? ¿Es portadora de conocimiento científico? ¿Sería digna de ser citada en un trabajo

[56] Ibíd, p.19
[57] Ibíd, p.19
[58] Ibíd, p.19
[59] Ibíd, p.19
[60] Ibíd, p.19
[61] Ibíd, p.19

académico como representativa del estado de conocimientos en la materia?[62]

1.6.9 Técnicas, algoritmos y métodos usados en web mining. El concepto web mining, agrupa a todas las técnicas, algoritmos y metodologías utilizadas para extraer información y conocimiento desde los web data, entre los cuales se cuentan[63,64].

- Redes neuronales artificiales: se trata de modelos predictivos no lineales que aprenden a través de la formación y se asemejan, estructuralmente hablando, a las redes neuronales biológicas.

- Self Organizing Feature Maps (SOFMs): Esta herramienta tiene una estructura semejante a las redes neuronales, pero en este caso el aprendizaje se da de manera competitiva, es decir, las neuronas compiten para ser activadas, y sólo lo hace una a la vez. La idea de este aprendizaje es que se compara un elemento con la red con el fin de encontrar la neurona más similar, o neurona ganadora. A partir de lo anterior se generar grupos de neuronas o clusters cuyas características son similares.

- K-Means: Este algoritmo se basa en la determinación de grupos o clusters dentro de un conjunto de datos. Para su funcionamiento se necesita como parámetro el número esperado de grupos (k). Cada uno de estos clusters estará representado por un centroide, que es el elemento cuyas características se parecen más a las de su conjunto (Obtenido mediante una medida de similitud). Este método tiene una alta performance, por lo que es posible repetirlo varias veces con distintos parámetros.

- Árboles de Decisión: Esta técnica se basa en la estimación de un resultado y toma de decisiones a partir de datos conocidos. La idea es identificar los atributos mínimos con los cuales se pueda deducir un resultado, clasificando los datos en una estructura de árbol y moviéndose a través de las ramas. Son conjuntos de decisiones, que generan reglas para la clasificación de un conjunto de datos, configurándose para ello en base a estructuras en forma de árbol[65].

- Support Vector Machines (SVMs): En comparación con las redes neuronales, tiene la ventaja de ser menos propensos al sobre aprendizaje, por lo tanto

[62] Ibíd, p.19
[63] MARKOV, Z. and LAROSE, D.T. Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage. John Wiley and Sons, New York, USA, 2007.
[64] VELÁSQUEZ, J.D. and PALADE, V. Op.cit.
[65] LOGICALIS. Modelos de data mining y las herramientas más usadas. Recuperado de Logicalis: Business and technology working as one, 2015: https://blog.es.logicalis.com/analytics/modelos-de-data-mining-y-las-herramientas-mas-usadas

pueden mantener un gran número de características y datos sin preocuparse de la complejidad del problema. La idea básica de esta herramienta es trabajar con ciertas funciones efectivas (Funciones de Kernel) que permitan tratar los datos a otro nivel dimensional y de esta forma trabajar con modelos complejos.

- Algoritmos Inspirados en la Vida. Se trata de una nueva familia de algoritmos cuya operación está basada en cómo ciertas especies, bacterias y la misma evolución con cambios genéticos, tratan de sobrevivir y perpetuarse en la vida.

- Algoritmos genéticos: esta herramienta, bastante utilizada para obtener modelos de data mining consiste en la aplicación de técnicas de optimización basadas en los conceptos de la combinación genética, la mutación y la selección natural[66].

1.7 METODOLOGÍA

1.7.1 Tipo de estudio. El estudio consistió en una investigación descriptiva de revisión sistemática.[67]

1.7.2 Fuentes de información. Se manejaron como fuentes de información las bases documentales Redalyc, Scielo, Scopus, IEEEXplore, Google Scholar y Web of Science.

1.8 DISEÑO METODOLÓGICO

Diferentes autores han diseñado metodologías para la realización de revisiones sistemáticas, principalmente orientadas al área de la salud. Sin embargo, dada la naturaleza del presente estudio se tomó la metodología propuesta por Kitchenham[68], en la cual se presenta un método para la realización de revisiones sistemáticas en el contexto de Ingeniería del Software, que involucra diferentes actividades independientes, razón por la cual fue elegida para ser aplicada en la investigación. Así, el protocolo de investigación del método propone tres fases fundamentales (Figura 1).

---

[66] Ibíd. p.2

[67] HURTADO, J. Metodología de la Investigación Holística. Caracas, Venezuela: Editorial SYPAL.

[68] KITCHENHAM, B. Procedures for performing systematic reviews (Joint Technical Report). Software Engineering Group, Department of Computer Science, Keele University and Empirical Software Engineering National ICT Australia Ltd. 2004.

La primera fase es la planificación de la revisión, la cual consiste en la preparación de los aspectos con los que se definirá la búsqueda de información. Consta de dos actividades principales:

(1) Identificación de la necesidad de la revisión

(2) Desarrollo de un protocolo de revisión.

La segunda fase es el desarrollo de la revisión que consiste propiamente en la búsqueda de la información, su identificación y síntesis. Consta de las siguientes actividades:

(1) Identificación de la investigación

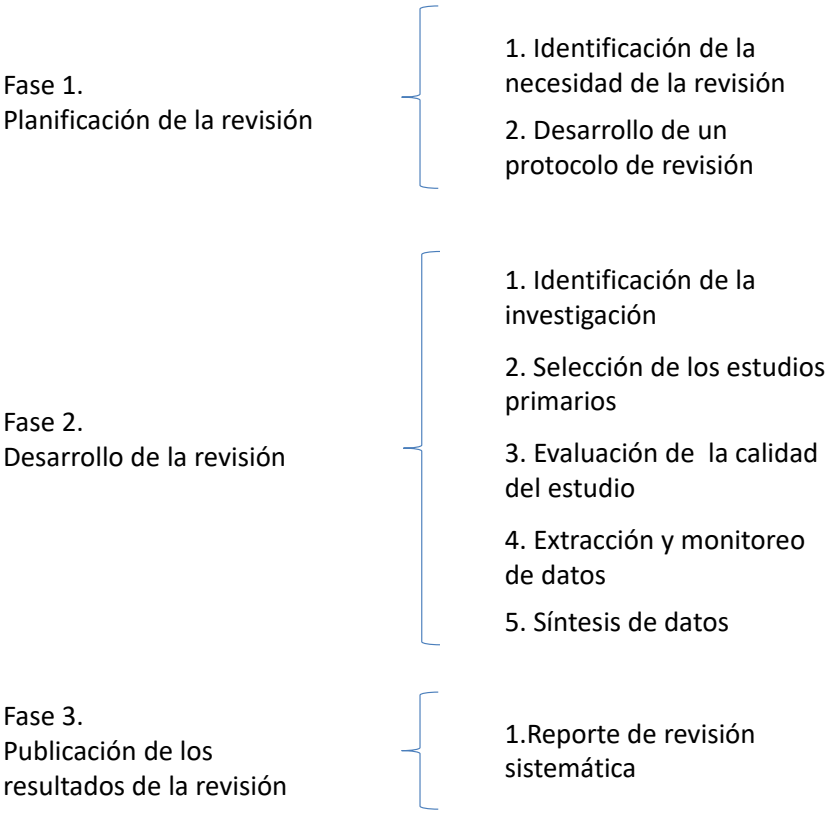
(2) Selección de los estudios primarios

(3) Evaluación de la calidad del estudio

(4) Extracción y monitoreo de datos

(5) Síntesis de datos

Figura 1. Protocolo de investigación

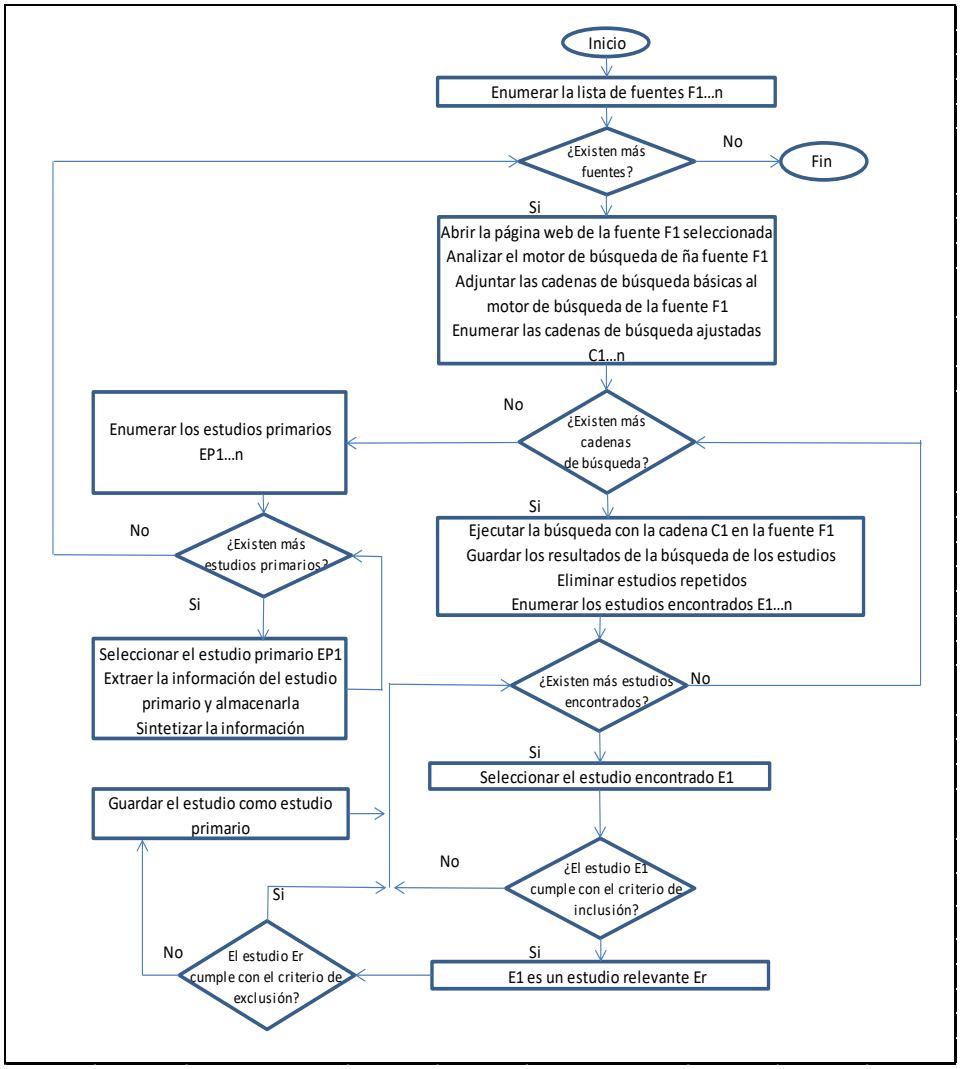


Fuente: Elaboración propia

La tercera fase es la publicación de los resultados de la revisión que consiste en la divulgación de los hallazgos en una revista arbitrada.
Para facilitar el desarrollo de las fases 1 y 2, planificación y ejecución de la revisión sistemática, se desarrolló una plantilla del protocolo para la revisión[69] indicada en la figura 2. El objetivo fue que sirviera como guía a los investigadores en Ingeniería del Software al conducir revisiones sistemáticas. En ella se ilustra el paso a paso que se siguió en cada parte de la revisión.

[69] BIOLCHINI, J. et al., Systematic Review in Software Engineering. Systems Engineering and Computer Science Department, UFRJ: Rio de Janeiro, Brazil, 2005.

Figura 2. Procedimiento para obtener los estudios primarios y sintetizar su información



Fuente: Extraído de Pino et al.[70]

1.8.1 Procedimiento. El desarrollo de la investigación fue llevado a cabo en cuatro etapas, así:

Etapa 1. Definición de los criterios de búsqueda de artículos sobre minería de contenido en la web a ser empleados en las bases documentales.

[70] PINO, F.J.; GARCÍA, F. & PIATTINI, M. Revisión sistemática de mejora de procesos software en micro, pequeñas y medianas empresas. Revista Española de Innovación, Calidad e Ingeniería del Software, 2(1):6-23, 2006.

En esta etapa se llevó a cabo la fase 1 del protocolo de investigación consistente en la planificación de la revisión, la cual constó de las siguientes actividades:

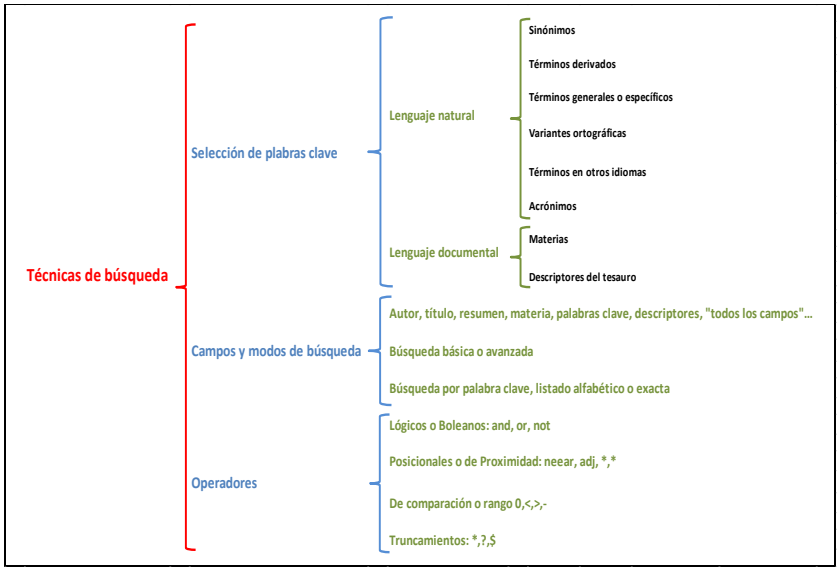
- Identificación de la necesidad de la revisión, la cual se enfocó en las principales técnicas que permiten realizar minería de contenido en la web.

- Desarrollo de un protocolo de revisión, el cual partió de la necesidad de la revisión identificada, luego se definió el nivel y la cobertura de la búsqueda, se seleccionaron las fuentes de información, se elaboró la estrategia de búsqueda (Figura 3), se efectuó la valoración de los resultados y se gestionó la información recuperada.

Etapa 2. Identificación de las técnicas que han sido empleadas para realizar minería de contenido en la web.

Esta etapa se llevó a cabo mediante el desarrollo de la fase 2 del protocolo de investigación, en la que fueron empleadas las técnicas de búsqueda indicadas en la figura 3. Una vez seleccionados los artículos científicos de las bases documentales se listaron, se efectuaron los respectivos RAE's, se llevó a cabo su análisis y se efectuó la extracción de la información acerca de las técnicas de minería de contenido web.

Figura 3. Técnicas de búsqueda de información



Fuente: Elaboración propia con base en García[71]

[71] GARCÍA, D.F. Revisión sistemática de literatura en los trabajos de final de Máster y en las tesis Doctorales. Grupo de investigación en InterAcción y eLearning (GRIAL). España: Universidad de Salamanca. 2017.

Para ello se tuvieron en cuenta las siguientes actividades:

- Identificación de la investigación: técnicas de minería web de contenido
- Selección de los estudios primarios: para lo cual se emplearon las técnicas de búsqueda de información ilustradas en la figura 3
- Evaluación de la calidad del estudio: para ello se tuvo en cuenta el cumplimiento de los criterios de inclusión
- Extracción y monitoreo de datos: esta actividad consistió en realizar el registro en la matriz RAE de autor, título, resumen, descriptores, entre otros
- Síntesis de datos: mediante esta actividad se obtuvo la información requerida sobre las técnicas de minería web de contenido, propósito de la investigación

Etapa 3. Caracterización de las principales técnicas para realizar minería de contenido en la web.

A partir de la identificación de las técnicas de minería de contenido web se procedió a realizar una descripción detallada de las mismas, con base en la información recopilada en la revisión sistemática.

Etapa 4. Elaboración del reporte de la revisión sistemática. Esta etapa se llevó a cabo con la fase 3 del protocolo de investigación consistente en realizar la publicación de los resultados en una revista arbitrada. Para ello se identificó la revista, se organizó la publicación acorde con el formato de la revista seleccionada y se envió a ser considerada por el comité editorial.

2. DEFINICIÓN DE LOS CRITERIOS DE BÚSQUEDA DE ARTÍCULOS SOBRE MINERÍA DE CONTENIDO EN LA WEB

A continuación, se describen los criterios empleados para la búsqueda de artículos sobre minería web en las bases documentales seleccionadas.

2.1 TÉRMINOS DE BÚSQUEDA

En el estudio fueron empleados los términos “Minería web”, “Contenido web” “Técnicas de minería web”; “Web mining", "Web content", "Web mining techniques” para delimitar la búsqueda a artículos que fueran de interés, además fue importante la utilización del boleano "and" ya que se necesitaba la relación directa existente entre la minería web y los contenidos web.

2.2 BASES DE DATOS

Para el desarrollo de la investigación se emplearon las bases de datos Redalyc, Scielo, Scopus, IEEEXplore, Google Scholar y Web of Science.

2.3 ALGORITMOS DE BÚSQUEDA

En Redalyc se emplearon los algoritmos de búsqueda:

- “Minería web” AND “Contenido web”
- “Técnicas de minería web”
- “Web mining" AND "Web content"
- "Web mining techniques”

En Scielo se emplearon los algoritmos de búsqueda:

- “Minería web” AND “Contenido web”
- “Técnicas de minería web”
- “Web mining" AND "Web content"
- "Web mining techniques”

En Scopus se emplearon los algoritmos de búsqueda:

- “Minería web” AND “Contenido web”
- “Técnicas de minería web”
- “Web mining" AND "Web content"
- "Web mining techniques”

En IEEEXplore se emplearon los algoritmos de búsqueda:

- “Minería web” AND “Contenido web”
- “Técnicas de minería web”
- “Web mining" AND "Web content"
- "Web mining techniques”

En Google Scholar se emplearon los algoritmos de búsqueda:
“Minería web” AND “Contenido web”

- “Técnicas de minería web”
- “Web mining" AND "Web content"
- "Web mining techniques”

En Web of Science se emplearon los algoritmos de búsqueda:

- “Minería web” AND “Contenido web”
- “Técnicas de minería web”
- “Web mining" AND "Web content"
- "Web mining techniques”

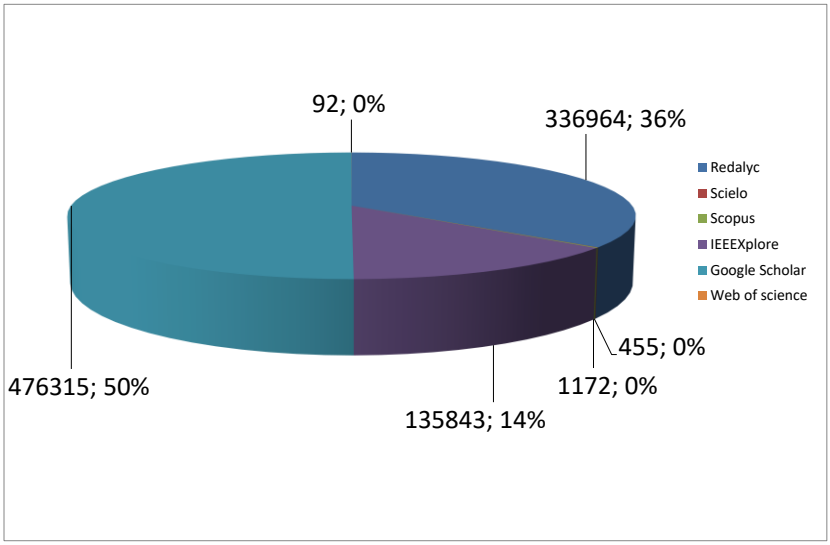
Con base en lo anterior, la búsqueda preliminar sin el empleo de filtros, produjo como resultado (Tabla 1), la existencia de 950.841 documentos correspondientes en un 50,10% a la base documental Google Scholar, en un 35,44% a la base documental Redalyc, y en un 14,29% a IEEEXplore, siendo el aporte de documentos de las bases documentales Scopus, Scielo y Web of science en conjunto inferior al 1% de los documentos existentes (Figura 4).

Tabla 1. Resultados de la búsqueda en las diferentes bases documentales de acuerdo con los algoritmos de búsqueda empleados sin el empleo de filtros

Base documental / algoritmo de búsqueda	"Minería web"	"Contenido web"	“Minería web” AND “Contenido web”	"Técnicas de minería web"	“Web mining" AND "Web content"	“Web mining"	"Web content"	"Web mining techniques”	TOTAL	% de documentos
Redalyc	31	250	336343	4	0	77	259	0	336964	35,44%
Scielo	9	65	2	0	3	62	314	0	455	0,04%
Scopus	0	63	0	0	7	130	907	65	1172	0,12%
IEEEXplore	0	0	110535	6721	2273	12371	1084	2859	135843	14,29%
Google Scholar	420	4490	57	48	9540	57700	401000	3060	476315	50,10%
Web of Science	3	6	4	8	0	0	71	0	92	0,01%
TOTAL	**463**	**4874**	**446941**	**6781**	**11823**	**70340**	**403635**	**5984**	**950841**	100,00%

Fuente: Elaboración propia (2018)

Figura 4. Distribución de documentos relacionados sin la aplicación de filtros

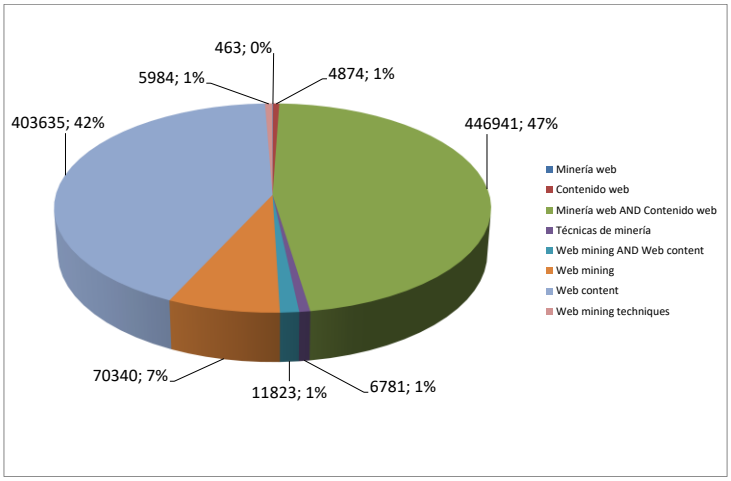


Fuente: Elaboración propia

Estos resultados se dan por el cubrimiento de las diferentes bases documentales donde en la actualidad Google scholar es quizá la de mayor espectro a nivel mundial.

Así mismo, al observar los términos que condujeron al mayor número de documentos se obtuvo que “Minería web” AND “Contenido web” generó el 47% de los resultados, "Web content" el 42,45% y “Web Mining” el 7,40% (Figura 5).

Figura 5. Resultados de la aplicación de los términos de búsqueda sin filtros



Fuente: Elaboración propia

Lo cual se explica porque la combinación de minería web y contenido web incluye los ítems de cada uno por separado, además de los que por su combinación se obtienen.

2.4 FILTROS

Se emplearon los siguientes filtros:

- Text: complete
- Publication year: 2014 - 2018

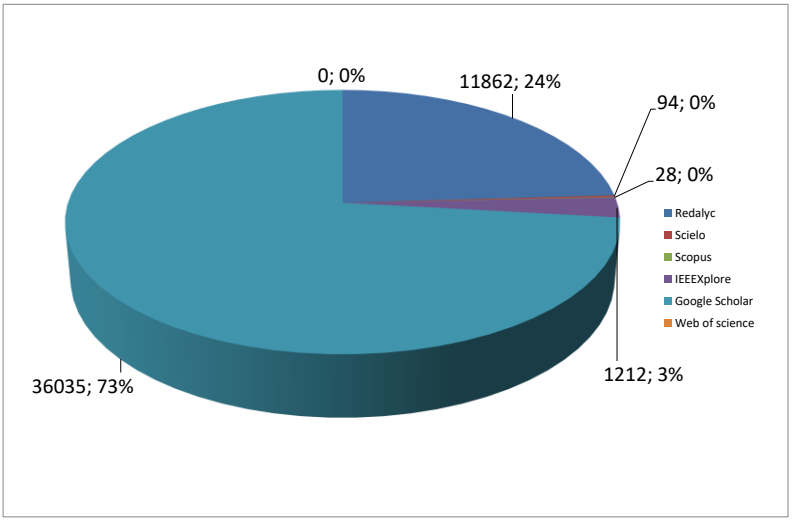
Al realizar la búsqueda empleando los dos filtros propuestos se obtuvieron los resultados de la tabla 2, en donde el número de documentos hallados se redujo a 49.231, es decir al 5,17% de los documentos existentes, en donde el 73,2% de ellos se encuentran en la base documental Google Scholar y el 24,09% en la base documental Redalyc. Tan solo el 2,71% se encuentran en las otras cuatro bases documentales revisadas (Figura 6). Así mismo, el término “web content” incluyó el 37,05% de los resultados obtenidos, “web mining” el 26,47% y “Minería web” AND “Contenido web” el 23,83% (Figura 7). A partir de lo anterior se puede evidenciar que tras el empleo de los filtros los términos mencionados se mantuvieron como los que mayor aporte de documentos generan aun cuando no con la misma importancia que sin su empleo.

Tabla 2. Resultados de la búsqueda en las diferentes bases documentales de acuerdo con los algoritmos de búsqueda empleados teniendo en cuenta el filtro de año 2014 – 2018

Base documental / algoritmo de búsqueda	"Minería web"	"Contenido web"	“Minería web” AND “Contenido web”	"Técnicas de minería web"	“Web mining" AND "Web content"	“Web mining"	"Web content"	"Web mining techniques”	TOTAL	% de documentos
Redalyc	9	81	11659	0	0	0	113	0	11862	24,09%
Scielo	1	5	0	0	2	4	82	0	94	0,19%
Scopus	0	0	0	0	0	6	19	3	28	0,06%
IEEEXplore	0	0	54	0	82	23	26	1027	1212	2,46%
Google Scholar	137	1990	21	13	2100	13000	18000	774	36035	73,20%
Web of Science	0	0	0	0	0	0	0	0	0	0,00%
TOTAL	**147**	**2076**	**11734**	**13**	**2184**	**13033**	**18240**	**1804**	**49231**	**100,00%**

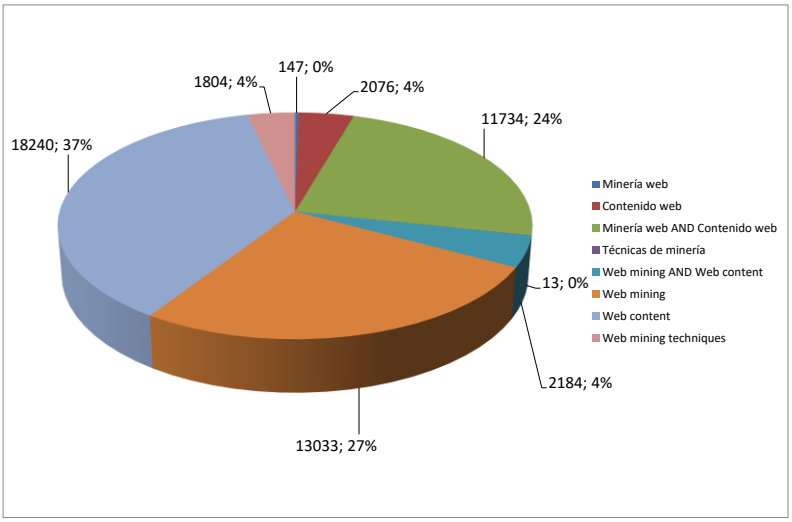
Fuente: Elaboración propia (2018)

Figura 6. Distribución de documentos relacionados con la aplicación de filtros



Fuente: Elaboración propia

Figura 7. Resultados de la aplicación de los términos de búsqueda con la aplicación de filtros



Fuente: Elaboración propia

2.5 CRITERIOS DE INCLUSIÓN

Como criterios de inclusión se buscaron artículos publicados durante el periodo 2014 - 2018, artículos completos ya sea en idioma inglés y/o español, que contemplen técnicas o el desarrollo de las mismas en minería web de contenido.

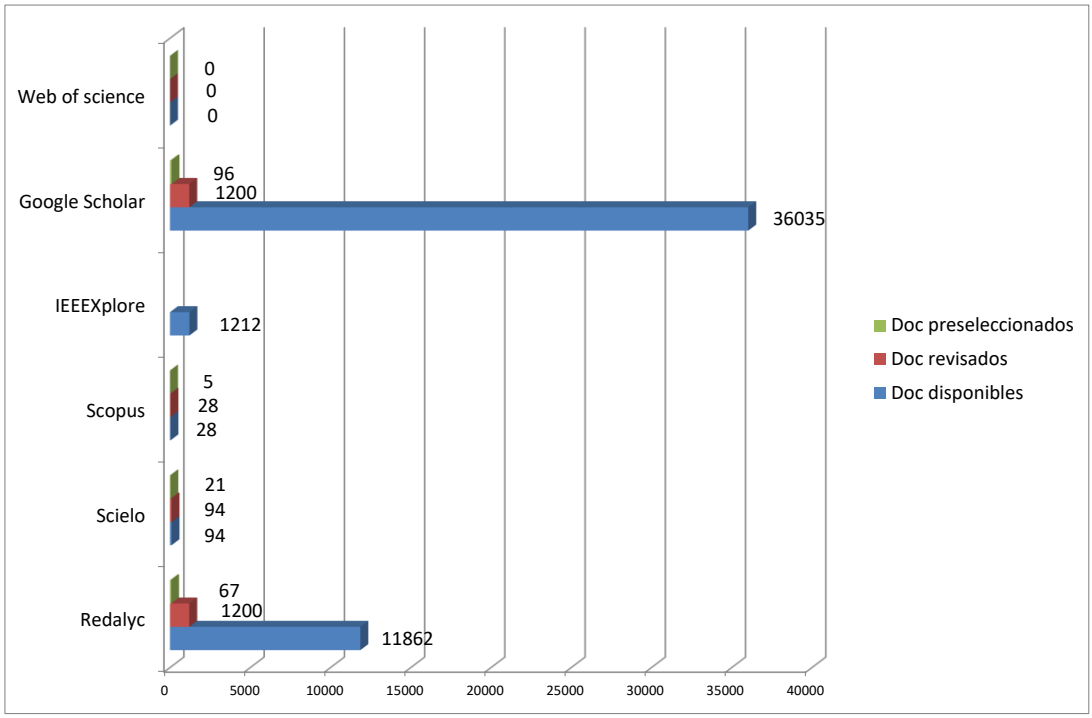
3. IDENTIFICACIÓN DE LAS TÉCNICAS QUE HAN SIDO EMPLEADAS PARA REALIZAR MINERÍA DE CONTENIDO EN LA WEB

3.1 PRESELECCIÓN DE ARTÍCULOS

Después de la revisión de 410 artículos de la base documental Redalyc se han preseleccionado 47. Así mismo, una vez revisados 94 artículos de la base documental Scielo se han preseleccionado 21. De la base documental Scopus se han revisado 28 artículos de los cuales se preseleccionaron 5. Finalmente, de la base documental Google Scholar se revisaron 352 artículos de los cuales se preseleccionaron 61 (Figura 8).

Fueron revisados 3734 artículos de los cuales se preseleccionaron 189, cuyos RAE's se registraron en una matriz (Fragmento mostrado en el Anexo 1).

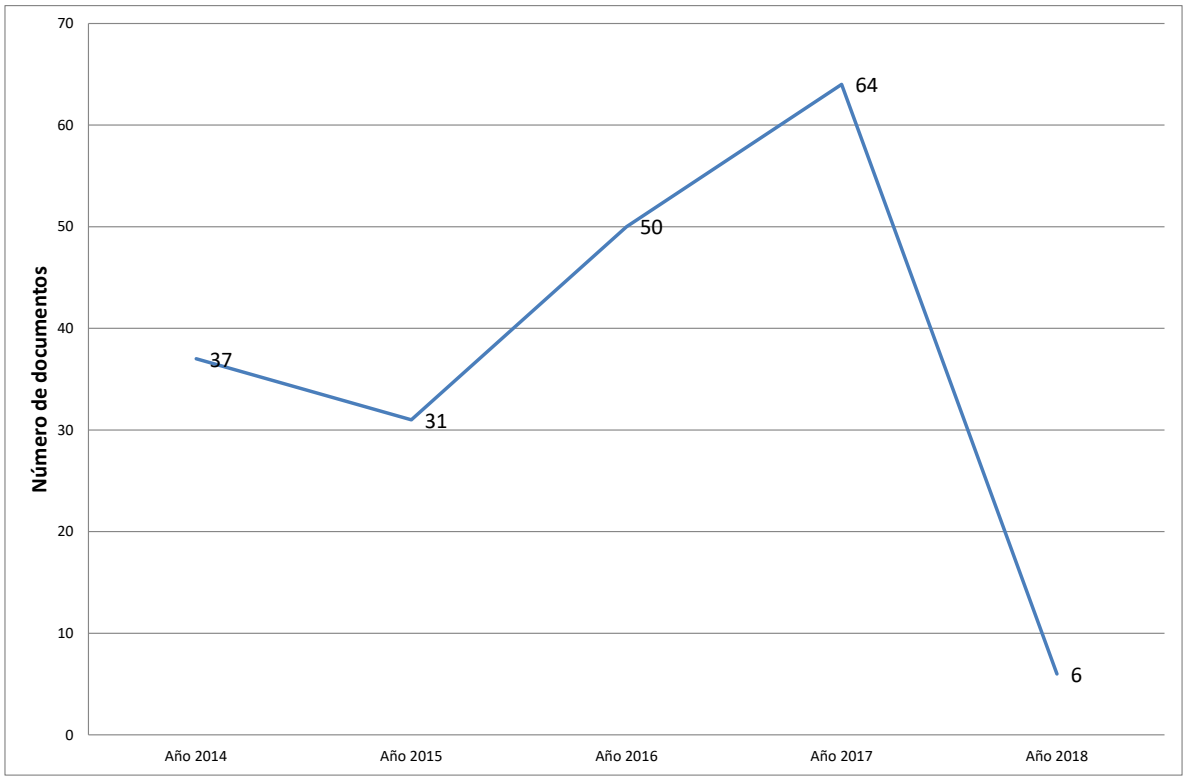
Figura 8. Relación documentos preseleccionados vs documentos disponibles por base documental



Fuente: Elaboración propia (2018)

Con base en lo anterior se preseleccionaron el siguiente número de artículos por año (Figura 9).

Figura 9. Número de documentos preseleccionados por año



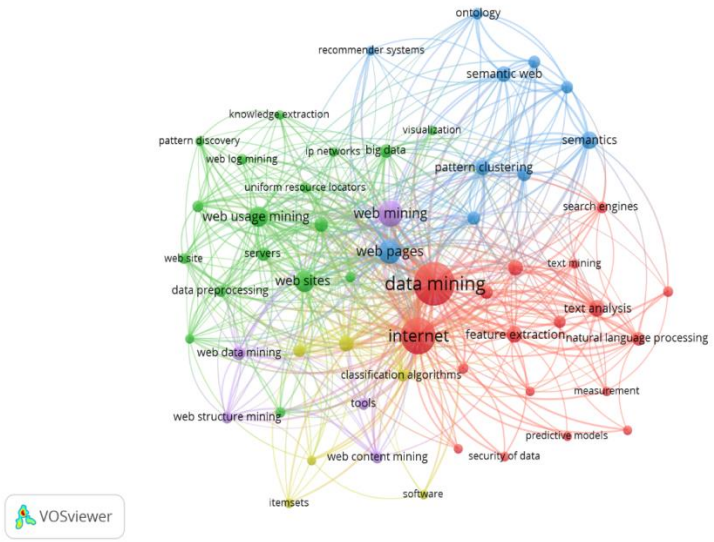
Fuente: Elaboración propia

La figura 9 indica que el año de mayor producción de documentos relacionados con la temática, de acuerdo con los resultados de la búsqueda fue el 2017, con el 33,86% y el año 2016 con el 26,45% de los documentos preseleccionados.

3.2 ANALISIS BIBLIOMÉTRICO

Del resultado obtenido de las diferentes bases de datos digitales utilizadas para el desarrollo de la revisión sistemática, se empleó el software de acceso libre VosViewer, con el fin de realizar el análisis de documentos y consolidación de los indicadores de acuerdo con la ecuación de búsqueda y así identificar varios indicadores, uno de ellos son los términos comunes dentro de los artículos consultados como se muestra en la figura 10 y figura 11.

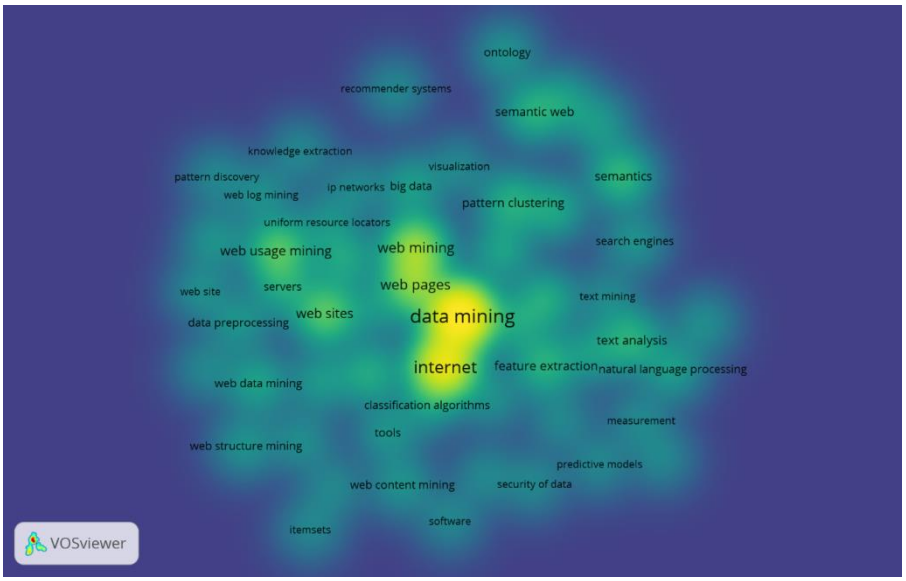
Figura 10. Términos más comunes en artículos



Fuente: Elaboración propia.

La figura 10 resalta los términos que se encuentran en los artículos, y que presentan una mayor cantidad de usabilidad. Por ejemplo, data mining, internet, web page, web mining. Entre tanto el circulo más grande visualizado en la gráfica, representa la palabra más frecuente que aparece en las publicaciones o artículos

Figura 11. Mapa de calor de términos

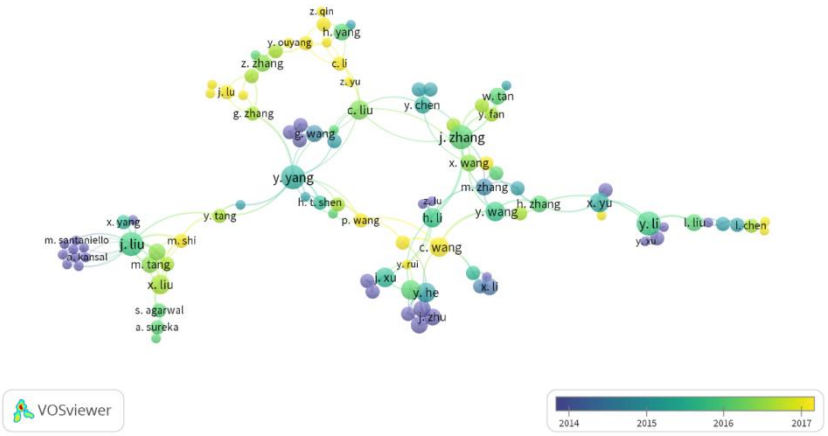


Fuente: Elaboración propia.

La figura 11 muestra la correlación y cercanía entre los artículos, donde el color amarillo indica que es más alta la cercanía entre dichos términos. El mapa revela bastantes subtemas relacionados con la revisión sistemática desarrollada.

De igual manera, por medio de este software se discriminan los autores que más artículos han escrito en los últimos cinco años, así como se muestra en la figura 12 y 13.

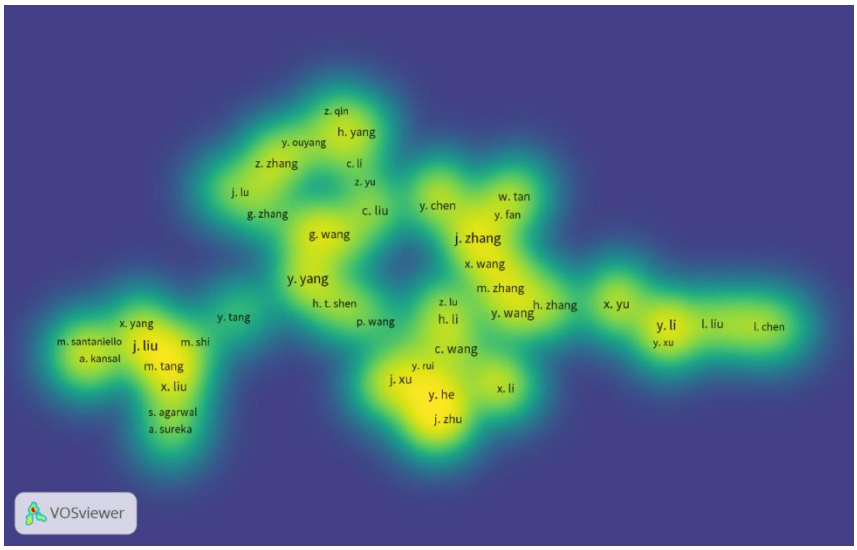
Figura 12. Autores de artículos



Fuente: Elaboración propia.

En la figura anterior, se identifica que en los años que más autores realizaron publicaciones fue en rango de 4 años, comprendidos entre 2014 y 2017, pero con una mayor tendencia en 2016 y 2017.

Figura 13. Mapa de calor: Autores



Fuente: Elaboración propia.

Para la construcción y visualización de las figuras 10, 11, 12 y 13, se realizó la exportación de los resultados obtenidos de las bases de datos mediante los términos de búsqueda, con un archivo de extensión RIS, el cual se coloca en el software que se encarga de generar estas gráficas.

3.3 TÉCNICAS EMPLEADAS PARA REALIZAR MINERÍA DE CONTENIDO WEB

A partir de los artículos seleccionados se encontró que la minería web se refiere al proceso general de descubrimiento de información potencialmente útil y previamente desconocida o conocida de los datos web. Así, la minería web se utiliza para capturar información relevante, creando nuevos conocimientos a partir de los datos, la personalización de la información, aprendizaje sobre los consumidores o usuarios individuales y muchos otros. La minería web utiliza técnicas de minería de datos para descubrir automáticamente y extraer información del World Wide Web[72].

La técnica tradicional de búsqueda en la web fue a través de los contenidos. La minería de contenido web es la extensión del trabajo realizado por motores de búsqueda. La minería de contenido web se refiere al descubrimiento de información útil desde los contenidos de la web como texto, imágenes, videos, etc. Así, existen dos enfoques empleados en la minería web de contenidos, a saber: un enfoque basado en agentes y un enfoque de base de datos[73].

[72] JOHNSON, F. & KUMAR, S. Web Content Mining Techniques: A Survey. International Journal of Computer Applications (0975 – 888), June 2014, vol 47, no.11, p. 44-50.
[73] Ibíd, p.45

El enfoque basado en agentes cuenta con tres tipos: Agentes de búsqueda inteligentes, Agente de filtrado de información / Categorización, Agentes de web personalizada. Los primeros, automáticamente buscan información de acuerdo con las características particulares del dominio y los perfiles de usuario. Los segundos, usan técnicas de filtrado de datos según instrucciones predefinidas. Y los terceros, aprenden las preferencias del usuario y descubren documentos relacionados con esos perfiles de usuario.

Por otra parte, el enfoque de la base de datos consiste en una base de datos bien formada que contiene esquemas y atributos con dominios definidos.

Sin embargo, la minería de contenidos web se complica cuando tiene que minar datos no estructurados, estructurados, semiestructurados y multimedia, los cuales han conducido al desarrollo de las técnicas de minería de contenido web que se ilustran en la figura 14.

Es por ello que se contemplan también las técnicas de minería de datos, las cuales se pueden clasificar como se muestra en la **¡Error! No se encuentra el origen de la referencia.**

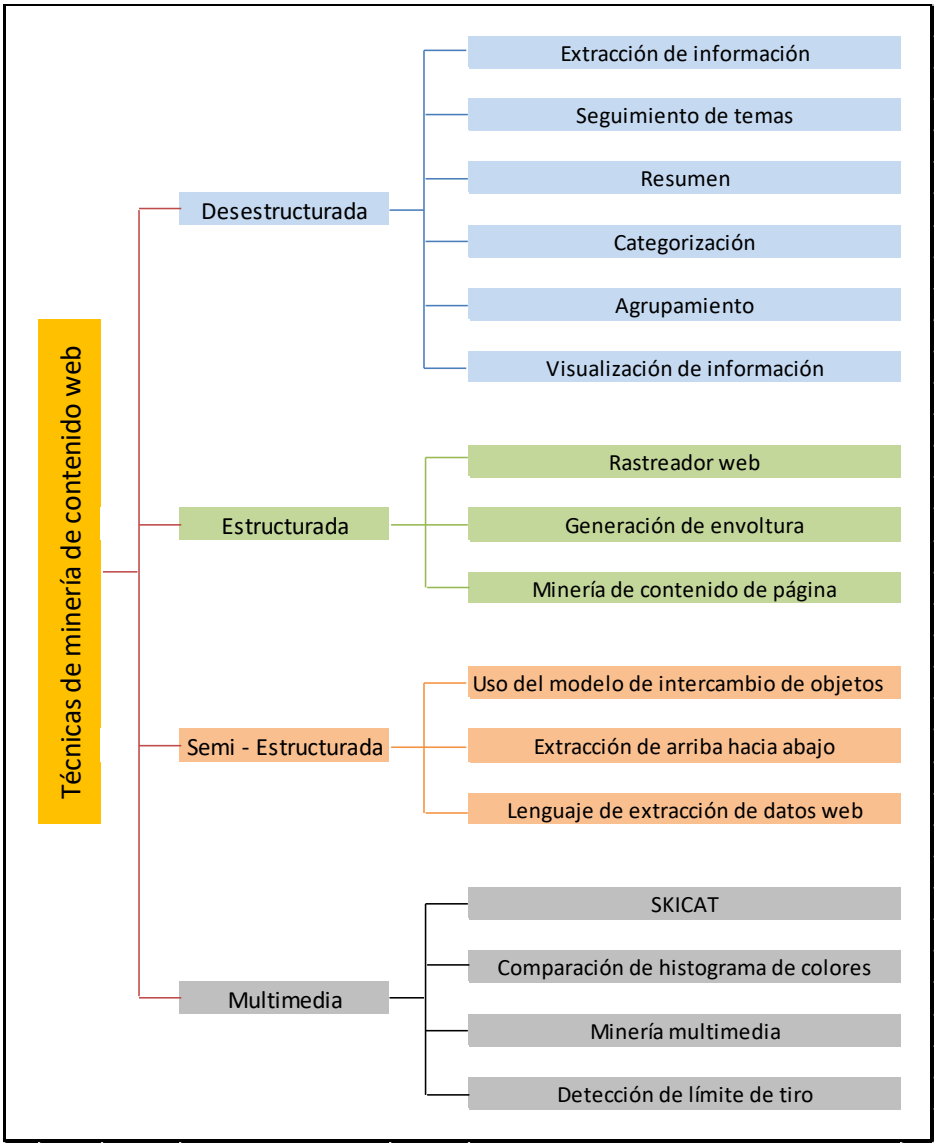
Según Sosa & Sosa[74] la primera técnica denominada “predictivas” se refiere a aquellas en las que las variables pueden clasificarse inicialmente en dependientes e independientes con base en un conocimiento teórico previo. Algunos algoritmos son los de tipo de regresión, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas.

Por otra parte, las “descriptivas” son aquellas en las que todas las variables tienen al inicio el mismo nivel o grado de pertenencia. Se crean automáticamente iniciando del reconocimiento de patrones. En este grupo se pueden encontrar técnicas de segmentación, agrupación (clustering), reducción de la dimensionalidad, entre otras.

Por último, “las auxiliares” son más limitadas y usadas de apoyo superficial. Se basan en técnicas de estadística descriptiva, consultas e informes dirigidas generalmente a la presentación y verificación.

[74] SOSA, M.O. & SOSA, E.C. Estudio de técnicas de Data Mining aplicadas al análisis de datos generados con la metodología Blended Learning. In XVI Workshop de Investigadores en Ciencias de la Computación. 2014.

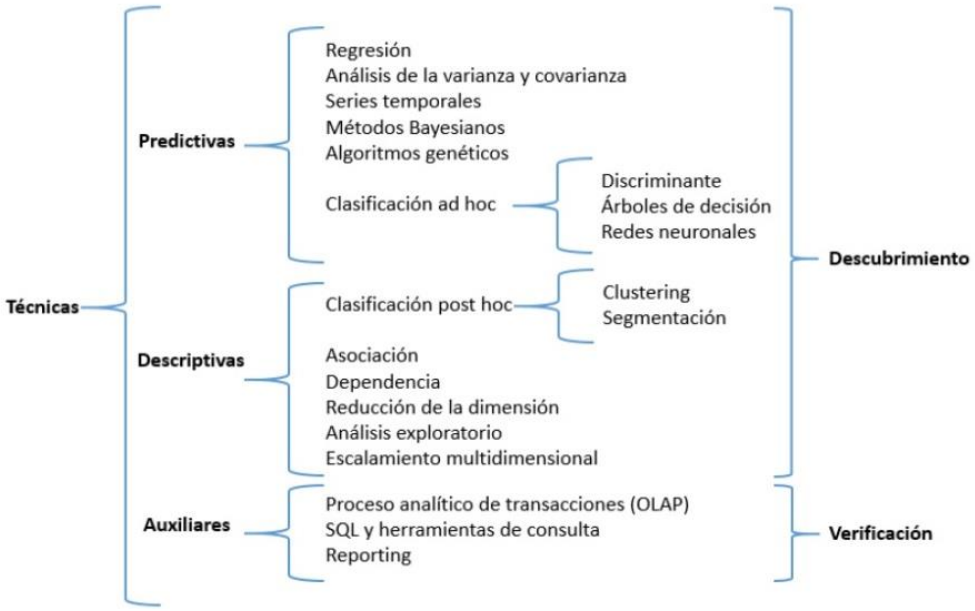
Figura 14. Técnicas de minería de contenidos web



Fuente: Elaboración propia con base en Johnson & Kumar[75]

[75] JOHNSON, F. & KUMAR, S. Op.Cit. p.46.

Figura 15. Clasificación de las técnicas de minería de datos



Fuente: Búsqueda de patrones en el comportamiento de los visitantes de la plataforma “Oferto” de la Cámara de Comercio de Armenia y del Quindío, a través de la aplicación de minería web[76].

[76] ANGARITA, D.A. & MUÑOZ, J.J. Búsqueda de patrones en el comportamiento de los visitantes de la plataforma “Oferto” de la Cámara de Comercio de Armenia y del Quindío, a través de la aplicación de minería web. Manizales: Universidad Autónoma de Manizales. 2016.

4. CARACTERIZACIÓN DE LAS PRINCIPALES TÉCNICAS QUE HAN SIDO EMPLEADAS PARA REALIZAR MINERÍA DE CONTENIDO EN LA WEB

4.1 MINERÍA WEB

La minería web consiste en la aplicación de técnicas de minería de datos a documentos y servicios de la web, es decir, en extraer la información, imágenes, textos, audio, video, documentos y multimedia de un sitio web. La minería web se diferencia de la minería de datos en que su procesamiento se realiza en línea, mientras que la otra se lleva a cabo fuera de línea[77].

Aunque estas técnicas procedan de la minería de datos, presentan sus propias características debido a la naturaleza de los datos presentes en la web, algunas características de estos datos son[78]:

- La mayor parte de los datos de la web tienen poca estructura (por ejemplo, tablas html) o casi ninguna (como pueden ser textos planos o PDFs).
- Los volúmenes de datos son muy altos y en algunos casos crecen de forma exponencial, con la problemática asociada (Big Data).
- Los datos (a nivel de páginas) están relacionados mediante links.
- Los datos tienen formatos muy variados como html, PDFs, imágenes, video, etc.
- Se mezclan datos fiables con otros de menor fiabilidad, dando lugar a inconsistencias.

Cuando un usuario utiliza la web para registros digitales que pueden ser las direcciones IP, navegador, sitios visitados, videos e imágenes descargadas que son almacenadas en los *logs* del servidor, estos son los que utiliza, precisamente, la minería web para el análisis y la obtención de conocimiento. Así, de acuerdo con los objetivos de análisis se puede dividir en tres tipos[79]:

- Minería de contenido web: la cual se encarga de extraer información o conocimiento del contenido de una página web. Con ella se puede descubrir descripciones de productos, fotografías publicadas, opiniones de clientes y sentimientos de consumidores.
- Minería de la estructura web: la cual se encarga de extraer información de los hipervínculos respecto a la estructura de la página que pudiera ser importante. Se puede descubrir comunidades de usuarios con intereses en común. Puede ser útil para clasificar o agrupar documentos.

---

[77] ARIAS, M.B. Minería de texto en medios sociales. Caso de estudio del proyecto Tranvía de Cuenca. Cuenca, Ecuador: Universidad del Azuay. 2016.

[78] ESPINOZA, Diego y ROJAS, Daniel. Minería de opiniones de usuarios en sitios web orientados al servicio. Trabajo de grado en Ingeniería civil informática. Valparaíso: Pontificia Universidad Católica de Valparaíso, Facultad de Ingeniería, Escuela de Ingeniería Informática. 2016. 189 p.

[79] ARIAS. Op cit., p.12

- Minería de los registros de navegación en la web: la cual se encarga de extraer información de la relación que existe entre documentos en la web por búsquedas anteriores, registrando búsquedas y accesos de los usuarios a los mismos, es decir, extrae información de los hábitos y preferencias de los usuarios.

Así mismo, de acuerdo con lo anotado por De la Calle[80], existe una parte especializada dentro de la minería de textos dedicada a la exploración de información en la web, que es la minería web o *web mining*. Esto se deriva de que la Internet se ha convertido en una fuente casi inagotable de información interesante para todo tipo de usuarios e investigadores provenientes de campos diversos como la biología, la medicina, la banca, los negocios o el marketing. El éxito de la Web 2,0 (redes sociales, blogs, noticias, etc.), ha dinamizado la generación de contenidos en internet haciéndolos crecer diariamente gracias a los aportes realizados por los usuarios de todo el mundo.

La Minería Web (Web Mining: WM), es un área que se desprendió del proceso de Minería de Datos y cuyo término fue acuñado por Oren Etzioni en 1996, y actualmente es un área de investigación extensa[81]. Algunos autores definen a la WM como el uso de técnicas para descubrir y extraer de forma automática información de los documentos y servicios de la web. La WM es el proceso de descubrir y analizar información útil de los documentos de la Web. Sin embargo, la Minería Web se puede definir como el descubrimiento y análisis de información relevante que involucra el uso de técnicas y acercamientos basados en la minería de datos (Data Mining: DM), orientados al descubrimiento y extracción automática de información de documentos y servicios de la Web, teniendo en consideración el comportamiento y preferencias del usuario. Entre los objetivos principales del WM se tiene:

- Descubrir recursos, extraer información, analizar datos e inferir generalidades.
- Obtener nuevos conocimientos provenientes de la información disponible en la Web.
- Encontrar información relevante.
- Optimizar el diseño y estructura del sitio web.

La Web, es una enorme colección de datos con información muy heterogénea, que posee un aumento en problemas de escalabilidad y dinamismo. Por consiguiente,

[80] DE LA CALLE, G. Modelo basado en técnicas de procesamiento de lenguaje natural para extraer y anotar información de publicaciones científicas. Madrid: Universidad Politécnica de Madrid. 2014.
[81] GUZMÁN, Daniel. Minería de datos sobre opiniones de clientes en agencias virtuales de alojamiento. Trabajo de grado en Ingeniería Civil Informática. Valparaíso: Pontificia Universidad Católica de Valparaíso, Facultad de Ingeniería, Escuela de Ingeniería Informática. 2016. 130 p.

la Web es un área fértil para la investigación de Minería Web, con la existencia de una enorme cantidad de información en línea.

Este proceso de Minería Web, se puede definir formalmente como “el proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web”[82]

Se suele usar la denominación WM para catalogar los tres tipos de actividades considerablemente diferentes mencionadas anteriormente. Todas estas actividades se enmarcan dentro de la MD y además, están relacionadas con la web; sin embargo, los datos que son objeto de la minería son diferentes.

4.2 MINERÍA DE CONTENIDO WEB

La Minería Web de contenido es la que clasifica y organiza los metadatos extraídos, con el objeto de recuperar y facilitar el acceso a la información. Tales metadatos pueden ser estudiados estáticamente mediante instantáneas de la web en un periodo determinado. La minería de datos consta de cuatro fases: recolección automática de la información importante para procesarla posteriormente; procesamiento de datos, ordenándolos y clasificándolos automáticamente; descubrimiento de patrones mediante hallazgo de frecuencias, reglas de asociación que permiten establecer estrategias de difusión de la información en las organizaciones; y análisis de patrones -se interpretan y validan los patrones[83].

La Minería de Contenido Web (WCM, Web Content Mining), se trata del descubrimiento de información útil de los contenidos, datos, documentos y servicios de la web. Sin embargo, los contenidos web no se componen únicamente de texto, sino también de audio, vídeo, datos simbólicos e hiperenlazados y metadatos[84]

De acuerdo con Campaña[85] La minería de contenido web a veces se llama la minería de textos web, porque el contenido del texto es la zona más ampliamente investigado. Las tecnologías que se utilizan normalmente en la minería de contenido web son PNL (procesamiento de lenguaje natural) e IR (recuperación de información).

La minería del contenido explora texto, imágenes, gráficos y video de una página web para determinar la relevancia del contenido. Esta exploración se efectúa después de la agrupación de las páginas web a través de la minería de estructura y

[82] Ibíd.,p16
[83] CORCHERO, P.; FERNÁNDEZ, M.R.; & HURTADO, M.A. Posicionamiento SEO del Centro de Investigación Flamenco Telethusa: Palabras Claves. Revista Centro Investigación Flamenco Telethusa,2016, 9(10), 22-29.
[84] GUZMÁN. Op.cit.
[85] CAMPAÑA, F.X. Aplicación de técnicas de Data Mining a bases de datos de contenido musical para identificar rasgos de personalidad de los usuarios en el Distrito Metropolitano de Quito. Sangolquí: Universidad de las Fuerzas Armadas. 2017.

proporciona los resultados en función del nivel de relevancia para una consulta sugerida. Con la enorme cantidad de información que está disponible en Internet, la minería de contenido proporciona las listas de resultados de los motores de búsqueda por orden de mayor relevancia relacionados a las palabras clave de la consulta.

La minería de texto se orienta hacia la información específica de los motores de búsqueda. Esto permite la exploración de toda la Web para recuperar contenido de grupos de exploración de páginas Web específicos. Los resultados son páginas transmitidas a los motores de búsqueda a través del más alto nivel de relevancia a la más baja. Sin embargo, los motores de búsqueda tienen la capacidad de proporcionar enlaces a páginas web por los miles de accesos de personas en relación con el contenido de búsqueda, este tipo de minería web permite la reducción de la información irrelevante.

La categorización de contenido web con una base de datos de contenido es la herramienta más importante para el uso eficiente de los motores de búsqueda. Un cliente que solicita información sobre un tema o artículo en particular lo contrario tendría que buscar a través de miles de resultados para encontrar la información más relevante para su consulta. Miles de resultados a través del uso de la minería de texto se reducen en este paso. Esto elimina la frustración y mejora la navegación de información en la Web.

Por ende, la capacidad para llevar a cabo la minería de contenido Web permite que los resultados de los motores de búsqueda maximicen el flujo de clics de clientes a un sitio web, o de determinadas páginas web del sitio, para ser visitada en numerosas ocasiones en relevancia a las consultas de búsqueda. La agrupación y organización de contenido Web en una base de datos de contenido permite una navegación eficaz de las páginas de los motores de búsqueda y los clientes. Imágenes, contenidos, formatos y la estructura Web se examinan para producir una mayor calidad de la información al usuario sobre la base de las solicitudes presentadas. Las empresas pueden maximizar el uso de la minería de textos para mejorar la comercialización de sus sitios, así como los productos que ofrecen.[86]

La minería de contenido web, también puede ofrecer resultados basados en datos numéricos por lo tanto se complementa a la minería de opinión aportando un análisis cuantitativo de los datos[87].

Es importante mencionar que la minería de contenido web se especializa en la localización de patrones en el texto de los documentos, y es en la cual se enfoca el desarrollo de este estudio. Debido a que como se mencionó anteriormente este tipo de minería aplica las técnicas de minería de datos en línea, se describirán a continuación su clasificación y en qué consiste cada una de ellas.

[86] Ibíd. p.20
[87] ESPINOSA y ROJAS. Op.cit.

4.3 ETAPAS DE LA MINERÍA WEB

Para poder procesar los datos y transformarlos en información útil, podemos distinguir una serie de etapas dentro del proceso global de la Minería Web.

- Selección y recopilación de datos: Lo primero es determinar qué es lo que se quiere obtener y cuáles son los datos que nos facilitarán esa información para lograr la meta. Posteriormente, se localizan los documentos o archivos a adquirir; capturándose y almacenándose los datos pertinentes. El objetivo de esta etapa es recuperar automáticamente los documentos más importantes, indexándolos para optimizar la búsqueda. El proceso de indexación es complejo, debido a la gran cantidad de páginas Web, además que estas cambian continuamente; por lo cual existen cuatro enfoques de indexación, los cuales son: indexación manual, automática, inteligente o basada en agentes y basada en Metadatos.

- Extracción y pre procesamiento de información: Se trata principalmente de filtrar y limpiar los datos recogidos. Una vez extraída la información determinada a partir de un documento (ya sea HTML, XML, TEXTO, PS, PDF, LaTeX, FAQs), se eliminarán los datos erróneos o incompletos, y se presentarán de manera ordenada, para luego realizar transformación de estos por medios automáticos. El objetivo es identificar y etiquetar el contenido esencial del documento, para mapear hacia algún modelo de datos. La extracción de la información entrega nueva información a partir de la estructura del documento y su representación.

- Minería: En esta etapa, se descubren automáticamente los modelos o patrones generales sobre un sitio Web, así como por múltiples sitios, utilizando recursos estadísticos, técnicas de Minería de Datos, etc.

- Análisis: Una vez teniendo los patrones identificados, es necesario interpretarlos; para esto, existen diversas herramientas que permiten entender, ya sea visualmente o por algún otro método que facilita la interpretación de dichos patrones.

4.4 TÉCNICAS DE MINERÍA DE CONTENIDOS WEB

A continuación, se describen las principales técnicas de minería de contenido web identificadas en la Figura 15.

4.4.1 Técnicas de minería de datos desestructurados. La minería de contenido se puede hacer con datos no estructurados tales como los textos. Donde la minería de datos desestructurados está dada por información desconocida. Así, la minería de texto es la extracción de información previamente desconocida procedente de diferentes fuentes de texto. La minería de contenido requiere la aplicación tanto de técnicas de minería de datos como de minería de texto. La minería de contenidos básicos es un tipo de minería de texto. Algunas de las técnicas usadas en minería de textos son la extracción de información, seguimiento de temas, resumen, categorización, agrupamiento y visualización de información[88].

- Extracción de información: consiste en la extracción de información de datos desestructurados mediante el uso de comparación de patrones. Para ello rastrea palabras clave y frases, y luego descubre la conexión de las palabras clave dentro del texto. Esta técnica es muy útil cuando hay un gran volumen de texto. Es la base de muchas de las técnicas usadas para minería de datos desestructurados. La extracción de información puede estar proporcionada por el módulo KDD porque la extracción de información tiende a transformar texto no estructurado a datos más estructurados. Primero la información se extrae de los datos y luego se usan diferentes tipos de reglas, para hallar la información perdida. Si las predicciones son incorrectas los datos son descartados.
- Seguimiento de temas: es una técnica en la que se verifican documentos vistos por el usuario y estudia los perfiles de éste. De acuerdo con cada usuario predice otros documentos relacionados con su interés. Ayuda a rastrear todas las historias posteriores en la corriente de noticias. La desventaja de esta técnica es que cuando se buscan temas determinados se puede obtener información que no está relacionada con el interés particular.
- Resumen: es una técnica empleada para reducir la longitud de los documentos manteniendo los puntos principales. Ayuda al usuario a decidir cuándo debe leer un tópico o no. En esta el tiempo tomado por la técnica para resumir el documento es menor que el tiempo empleado para leer el primer párrafo. El desafío en esta técnica es enseñar al software a analizar la semántica e interpretar el significado.
- Categorización: es la técnica para identificar los temas principales al colocar los documentos en un conjunto predefinido de grupos. Cuenta el número de palabras en un documento. No procesa la información actual. Decide el tópico principal desde el conteo. Ranquea el documento acorde con los

[88] JOHNSON, F. & KUMAR, S. Op.Cit. p.46.

tópicos. Así, los documentos que tienen la mayoría de su contenido dentro de un tópico particular son ranqueados de primeros.

- Agrupamiento: es una técnica empleada para agrupar documentos similares. Ayuda al usuario a seleccionar fácilmente los tópicos de interés. Esta técnica es principalmente empleada en la administración de sistemas de información.
- Visualización de información: utiliza la extracción de características e indexación de términos clave para construir una representación gráfica. A través de esta técnica los documentos que tienen similitudes son descubiertos. Ayuda a los usuarios a analizar visualmente los contenidos. Se usa principalmente para encontrar tópicos relacionados en grandes cantidades de documentos.

4.4.2 Técnicas de minería de datos estructurados. Las técnicas utilizadas para extraer datos estructurados son Rastreador web, Generación de envoltura y Minería de contenido de página.

- Rastreador web: hay dos tipos de rastreadores web, los cuales son llamados como rastreadores de web internos y externos. Los rastreadores son programas computarizados que atraviesan la estructura del hipertexto en la web. Los rastreadores externos rastrean a través de sitios desconocidos. Los internos a través de páginas internas de los sitios web los cuales son retornados por los rastreadores externos.
- Generación de envoltura: provee información sobre la capacidad de las fuentes. Las páginas web son ranqueados por los motores de búsqueda tradicionales. De acuerdo con las páginas web consultadas la recuperación se hace utilizando el valor de rango de la página. Las fuentes son las que responden a la consulta y a los tipos de salida.
- Minería de contenido de página: es una técnica de extracción de datos estructurados que funciona en las páginas clasificadas por motores de búsqueda tradicional.

4.4.3 Técnicas de minería de datos semi-estructurados. Las técnicas utilizadas para la minería de datos semiestructurada son el Modelo de intercambio de objetos (OEM), extracción de arriba hacia abajo y el lenguaje de extracción de datos web.

- Uso del modelo de intercambio de objetos: La información relevante se extrae de datos semiestructurados y están integrados en un grupo de información útil y almacenados en el modelo de intercambio de objetos (OEM). Ayuda al usuario a entender la estructura de información en la web más precisamente. Es el más adecuado para un ambiente heterogéneo y dinámico. Una característica principal del modelo de intercambio de objetos es el que lo describe a sí mismo, ya que no hay necesidad de describir de antemano la estructura de un objeto.

- Extracción de arriba hacia abajo: En la extracción de arriba hacia abajo, extrae objetos complejos de un conjunto de fuentes web ricas y se convierte en objetos menos complejos hasta que los objetos atómicos hayan sido extraídos.
- Lenguaje de extracción de datos web: convierte los datos web a datos estructurados y los entrega a los usuarios finales. Almacena datos en forma de tablas.

4.4.4 Técnicas de minería de datos de multimedia. Algunas de las Técnicas de Minería de Datos Multimedia son SKICAT, Comparación de histograma de colores, Minería multimedia y Detección de límite de tiro.

- SKICAT: Utiliza la técnica de aprendizaje automático para convertir objetos de clases utilizables para los humanos. Integra una técnica para procesamiento de imagen y clasificación de datos que ayuda a clasificar un conjunto de clasificación muy grande.
- Comparación de histograma de colores: consiste en la ecualización y suavizado de un histograma de color. La ecualización intenta descubrir correlación entre los componentes de color. El problema que enfrenta la ecualización es un problema de datos dispersos que es la presencia de artefactos no deseados en imágenes ecualizadas. Este problema es resuelto mediante el uso de suavizado.
- Minería multimedia: Consta de cuatro pasos principales. Imagen excavadora para extracción de imágenes y de video, un preprocesador para la extracción de las características de la imagen y se almacenan en una base de datos, un núcleo de búsqueda se utiliza para hacer coincidir consultas con imagen y video disponibles en la base de datos. El módulo descubierto realiza rutinas de minería de información de imágenes para rastrear los patrones en imágenes.
- Detección de límite de tiro: Es una técnica en la cual los límites son automáticamente detectados entre tomas en video.

4.5 TÉCNICAS DE MINERÍA DE DATOS

A continuación, se describen las principales técnicas de minería de datos predictivas:

- Arboles de decisión: los algoritmos de árbol de decisión consisten en organizar los datos en elecciones que compiten formando ramas de influencia después de una decisión inicial. El tronco del árbol representa la decisión inicial, y empieza con una pregunta de sí o no, como tomar o no

desayuno. Tomar desayuno y no tomar desayuno serían las dos ramas divergentes del árbol, y cada elección posterior, tendría sus propias ramas divergentes que llevan a un punto final.

- El algoritmo K-means: se basa en el análisis de grupos. Trata de dividir los datos recogidos en bloques –clústers separados y agrupados por características comunes.

- Máquinas de vectores de soporte: toman datos de entrada y predicen cuál de las dos posibles categorías incluye los datos de entrada. Un ejemplo sería recoger los códigos postales de un grupo de votantes e intentar predecir si un votante es demócrata o republicano.

- El algoritmo Apriori: normalmente controla los datos de transacciones. Por ejemplo, en una tienda de ropa, el algoritmo podría controlar qué camisas suelen comprar juntas los clientes.

- El algoritmo EM: define un parámetro analizando los datos, y predice la posibilidad de una salida futura o evento aleatorio dentro de los parámetros de datos. Por ejemplo, el algoritmo EM podría intentar predecir el momento de una siguiente erupción de un géiser según los datos de tiempo de erupciones pasadas.

- Algoritmo PageRank: es un algoritmo base para los motores de búsqueda. Puntúa y estima la relevancia de un trozo determinado de datos dentro de un gran conjunto, como un único sitio web dentro de un conjunto mayor de todos los sitios web de Internet.

- Algoritmo AdaBoost: funciona dentro de otros algoritmos de aprendizaje que anticipan un comportamiento según los datos observados para que sean sensibles a extremos estadísticos. Aunque el algoritmo EM puede sesgarse debido a un géiser que tiene dos erupciones en menos de un minuto, cuando normalmente tiene erupción una vez al día, el algoritmo AdaBoost modificaría la salida del algoritmo EM analizando la relevancia del extremo.

- Algoritmo del vecino k más cercano: reconoce patrones en la ubicación de los datos y los asocia con un identificador mayor. Por ejemplo, si se quiere asignar una oficina postal a cada ubicación geográfica del hogar y se tiene un conjunto de datos para cada ubicación geográfica del hogar, el algoritmo del vecino k más cercano asignará las casas a la oficina postal más cercana según su proximidad.

- Naive Bayes: predice la salida de una identidad basándose en los datos de observaciones conocidas. Por ejemplo, si una persona tiene una altura de 6

pies y 6 pulgadas (1,97 m) y tiene una talla 14 de zapatos, el algoritmo Naive Bayes podría predecir con una determinada probabilidad que la persona es un hombre.

- Algoritmo Classification and Regressive Tree (CART): al igual que los análisis de árboles de decisión, organiza los datos según opciones que compiten, como si una persona sobrevive a un terremoto. Al contrario de los algoritmos de árboles de decisión, que sólo pueden clasificar una salida o una salida numérica basada en regresión, el algoritmo CART puede usar las dos para predecir la probabilidad de un evento.

Para efectos de los tomadores de decisiones en la práctica, poco importa qué técnica se empleó para llegar a una y otra conclusión. Lo relevante para ellos es la veracidad, probabilidad y calidad de las conclusiones obtenidas en el análisis de un escenario o modelo determinado de datos para una situación particular. Lo que sí es claro es que estas técnicas y algoritmos ya están probados y validados en multitud de escenarios que comprueban y verifican su exactitud. De otra parte, es importante recalcar que son de vital importancia los datos y su calidad, y las herramientas informáticas –Softwareque implementan estas técnicas y algoritmos[89].

A continuación, se describen las principales técnicas de minería de datos descriptivas:

- Clustering de documentos. Puede definirse como la tarea de separar documentos en grupos. El criterio de agrupamiento se basada en las similitudes existentes entre ellos. Sus aplicaciones más importantes son mejorar el rendimiento de los motores de búsqueda de información mediante la categorización previa de todos los documentos disponibles, facilitar la revisión de resultados por parte del usuario final, agrupando los resultados tras realizar búsquedas. Por ejemplo, la sección de noticias de Google.

- Clustering conceptual. Reside en localizar todas las regularidades de un conjunto de grafos conceptuales en una jerarquía, para facilitar la navegación a través del grafo. Algunas formas más habituales suelen ser métodos como c-mean y métodos no tradicionales basados en redes neuronales del tipo Kohonen.

- Reglas de Asociación. Consiste en identificar relaciones de asociación o correlación entre un conjunto extenso de datos. Originalmente, las reglas de asociación surgen de la necesidad de muchas industrias de encontrar relaciones entre los registros o transacciones almacenados en sus bases de datos. Las reglas de asociación fueron propuestas por Agrawal. Esta técnica

[89] GUTIÉRREZ, Jahir y MOLINA, Bernardo. Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. En: ONTARE: Revista de investigación de la Facultad de Ingeniería, 2015, vol.3, no.2, p.33-51.

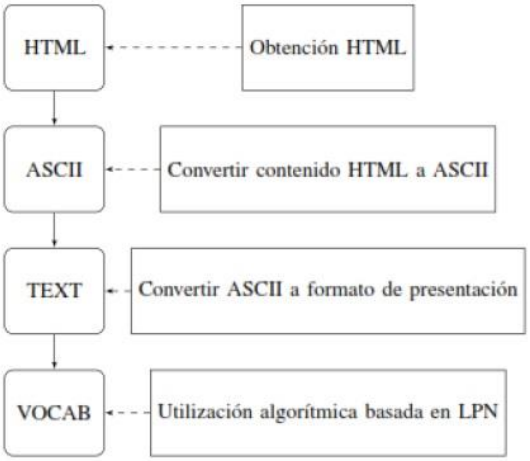
> fue desarrollada específicamente para tareas de minería de datos, en ella se intenta encontrar patrones en forma de reglas de tipo “IF-THEN” en un conjunto de ítems frecuentes. Por ejemplo, en un contexto de reportes de ventas de un supermercado, una regla de tipo” podría interpretarse como “las personas que compran cervezas también compran pañales”[90].

La minería de datos o exploración de datos (es la etapa de análisis de "Knowledge Discovery in Databases" o KDD) es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos[91]. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases, gestión y procesamiento de datos, se tienen en cuenta los aspectos del modelo e inferencias, de métricas de intereses, de consideraciones de la Teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea[92].

Al respecto Barberán, Rivadeneira & Larrea[93], anotan que emplearon la minería web para recuperar información mediante la extracción de un conjunto de documentos XML que son generados para cada usuario de la red. Previamente la minería web se basa en la clasificación del conjunto de documentos por lo general de acuerdo con sus puntuaciones de relevancia. En la Figura 16 se muestra el proceso de captura y transformación de este tipo de información.

[90] ZABALA, F.R. Buscador de artículos científicos aplicando minería de datos. 2014. Instituto Tecnológico de la Paz.
[91] CAMPAÑA, F.X. 2017. Op.cit.
[92] Ibíd.
[93] BARBERÁN, J.; RIVADENEIRA, F. & LARREA, J. Análisis de comentarios y revisiones de usuario usando redes neuronales recurrentes y procesamiento de lenguajes naturales. Revista Informática y Sistemas, 2018, 2(1): 1-13.

Figura 16. Recuperación de contenido HTML



Fuente: Análisis de comentarios y revisiones de usuario usando redes neuronales recurrentes y procesamiento de lenguajes naturales[94]

Camana[95] por su parte, exploró las posibilidades de uso de minería de datos en el Ecuador. Para llevar a cabo el proceso de minería de datos, se necesita de fases, que permitan descubrir patrones interesantes y potencialmente útiles de información. Por ello existe la extracción del conocimiento, está principalmente relacionado con el proceso de descubrimiento conocido como Knowledge Discovery in Databases (KDD), se refiere al proceso no-trivial de descubrir conocimiento e información, potencialmente útil dentro de los datos contenidos en algún repositorio de información. En la figura 17, se muestran las etapas del KDD.

[94] Ibíd. p.2.
[95] CAMANA, Roberto. Potenciales aplicaciones de la minería de datos en Ecuador. En: Revista Tecnológica ESPOL – RTE, Julio 2016, vol. 29, no. 1, p.170-183.

Figura 17. Proceso KDD

Fuente: Imagen tomada de Camana[96]

A continuación, se explica cada una de las etapas del proceso.

- Selección de datos: Recopilación de datos relevantes y tipo de información obtenida de diferentes fuentes, para ser utilizadas en el preprocesado.
- Preprocesado: Consiste en una exploración, ya que, al venir de diferentes fuentes de datos, es necesario una limpieza, es decir; eliminar o corregir datos incorrectos, necesaria para la siguiente etapa.
- Transformación: Esta etapa consiste en la trasformación de los datos, es decir, la creación de nuevas variables a partir de las ya existentes y la normalización de datos preparados para su posterior análisis.
- Minería de datos: Consiste en la búsqueda de patrones de interés, en una determinada forma de representación, en función al problema a solucionar.
- Interpretación y evaluación: Se evalúan patrones que serán analizados por expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración.

[96] Ibíd. p.175.

El rápido crecimiento de datos y las necesidades de convertir en información útil, ha permitido a centros de investigación y universidades, utilicen datos históricos almacenados, y que estos aporten al conocimiento, en la toma de decisiones.

- Intereses de investigadores, para aplicar la minería de datos, en campos poco o nada explotados.
- Saturación de investigaciones, en un mismo campo de estudio.
- Aparición de nuevos campos de aplicaciones de la minería de datos.
- Investigaciones previas en un mismo campo, sin continuos procesos de investigación.

Para confirmar experimentalmente, la utilidad de la minería de datos, se puede dar dentro de los siguientes aspectos:

- Sistemas parcialmente desconocidos: Si el modelo del sistema que produce los datos, es bien conocido, entonces no necesitamos de la minería de datos ya que todas las variables son de alguna manera predecibles.
- Enorme cantidad de datos: Al contar con mucha información, en algunas bases de datos es importante para una empresa encontrar la forma de analizar "montañas" de información (lo que para un humano sería imposible) y que ello le produzca algún tipo de beneficio.
- Potente hardware y software: Muchas de las herramientas presentes en la minería de datos están basadas en el uso intensivo de la computación, un software eficiente, aumentará el desempeño del proceso de buscar y analizar información, algo humanamente imposible.

Al revisar las investigaciones acerca de la temática de minería de texto y minería web, se observa que en un 70% de los visitantes que navegan en la web cumplen con patrones de costumbre de visitar el sitio web varias veces. Esto genera que los patrones de asociación, clasificación y agrupamiento de clientes o de información, facilitara la gestión de la información para actividades futuras dentro de las empresas. Además, esto ayuda a validar que tipos de visitantes navegan por el sitio web dado, que tipo de usuarios prefieren un contenido específico, los rasgos o características tiene en común, si son fieles al sitio web. También se observó las tendencias que los usuarios prefieren de acuerdo al tipo de contenido y estructura, todo lo expuesto anteriormente estos datos depende del tipo de aplicación dado por las herramientas text mining o web mining[97].

En muchas áreas las aplicaciones de minería de texto y minería web son muy utilizadas con el objetivo de mejorar la gestión de la información en la www. En el

[97] GUEVARA, Gladys; GUEVARA, Cristian y ELIZONDO, Daniel. Tratamiento de la información en la web: Text Mining y Web Mining. En: Revista Científica de Investigación actualización del mundo de las Ciencias, octubre, 2017, vol. 1, no. 4, p. 403-418.

área empresarial estas aplicaciones son de gran ayuda ya que permiten direccionar información específica a departamentos que lo soliciten. Como por ejemplo cuando llega un email enviado por un proveedor este lo redirecciona al departamento de compras, dependiendo del patrón de búsqueda que se haya aplicado para identificar el contenido y validarlo. Otra área en la que está siendo utilizada esta herramienta es la de mercados en la Web, extrae el conocimiento sobre estadísticas de manejo de determinados conceptos y temas en la web. En el área Bibliotecaria extrae la información más relevante del documento, compara la información automáticamente en los grupos de documentos con temas a fines, mostrándolos en la web de manera indexada[98].

Al realizar una comparación sobre la aplicación del text mining en la recuperación de la información en la web de acuerdo con las posiciones de los autores mencionados, se observaron coincidencias relacionadas con: Extracción de información, clasificación de la información o documentos, extracción de conocimientos, y elaboración de resúmenes.

[98] Ibíd. p.416

5. CONCLUSIONES

La minería web brinda herramientas muy útiles para el análisis de los contenidos web y de los datos, textos, documentos, imágenes, hiperenlaces, entre otros dentro de la web, una vez identificados los patrones de comportamiento que ayuden a la toma de decisiones permitirá establecer en cada empresa, el tipo de técnica que va a utilizar de acuerdo a las características, especificaciones, y problemas de ellas.

En esta revisión sistemática se evidenció que pese al volumen de documentos existentes sobre la temática (950.841), una vez empleados los filtros en la búsqueda se obtuvieron tan solo 49.231 equivalentes al 5,17% de los inicialmente estimados. Luego se llevó a cabo una revisión de 3734 documentos, equivalente al 7,58% de los identificados con filtros. De estos se seleccionaron 189 equivalentes al 5,06% de los revisados. Con lo que se concluye que aun cuando existe mucha información sobre una temática determinada, la que realmente sirve para los propósitos de la búsqueda es mínima.

Con esta revisión sistemática se logró identificar que la minería web de contenido es absolutamente necesaria en diversos campos de desempeño humano. Para ello se cuenta con técnicas propiamente dichas para minería de contenido web que se complementan con técnicas de minería de datos.

Así mismo, las técnicas identificadas y descritas ponen de manifiesto un procedimiento con el que se pueden aplicar en cualquier campo del conocimiento, en donde gracias a la disposición la minería web, se puede conocer, entender y predecir las conductas, rasgos, necesidades de los usuarios en la web, a través de esto se puede personalizar los procesos de búsqueda, agrupación de contenidos, aplicación de algoritmos, de acuerdo a los requerimientos.

6. RECOMENDACIONES

La minería web de contenido es un amplio campo de desarrollo de conocimiento que brinda herramientas muy útiles susceptibles de adecuaciones y especificaciones para la búsqueda de contenidos en la web en la actualidad. Cada sector de desarrollo humano tiene requerimientos específicos de información, en donde la minería web se convierte en la herramienta fundamental para la identificación de contenidos específicos dentro de grandes volúmenes de información disponible. Es por ello que este campo de estudio requiere ser profundizado mediante ejercicios de investigación orientados a aprovechar la información circulante en medios como las redes sociales y con los usuarios mediante sus perfiles de tal manera que la frecuencia de uso de ciertos contenidos conduzca a los sitios web de forma más ágil y puedan complementar las técnicas desarrolladas hasta el momento.

BIBLIOGRAFÍA

ABAD, José. Extracción de conocimientos de la diabetes tipo 1 utilizando la metodología de "Data mining". Tesis de grado Ingeniería Electrónica de Comunicaciones. Madrid: Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación. Departamento de Ingeniería de Telemática y Electrónica. 2015. 189p.

ANGARITA, David y MUÑOZ, Juan. Búsqueda de patrones en el comportamiento de los visitantes de la plataforma “Oferto” de la Cámara de Comercio de Armenia y del Quindío, a través de la aplicación de minería web. Tesis de Magíster en Gestión y Desarrollo de proyectos de Software. Manizales: Universidad Autónoma de Manizales. Facultad de Ingenierías. 2016. 160p.

ANGULO, Gerardo y CHARRIS, Maryuris. Desarrollo e implementación de una aplicación basada en minería de textos para la extracción y formateo de información de la base de datos de la USPTO que permita conocer tendencias tecnológicas. En: ResearchGate, Conference Paper, November, 2009, p.1-16

ARCILA, Carlos; BARBOSA, Eduar y CABEZUELO, Francisco. Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. En: El profesional de la información. 2016. vol. 25, no. 4, p. 623-631.

ARIAS, A.; MATOS, Y.; HEREDIA, J. y HEREDIA, D. Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación. En: Investigación y Desarrollo en TIC. 2016. vol. 7, no. 1, p. 14-20.

ARIAS, María Belén. Minería de texto en medios sociales. Caso de estudio del proyecto Tranvía de Cuenca. Tesis de grado de Ingeniero de Sistemas y Telemática. Cuenca: Universidad del Azuay. Facultad de Ciencias de la Administración. Escuela de Ingeniería de Sistemas y Telemática. 2016. 90 p.

ASENSIO, Elena. Aplicación de técnicas de minería de datos en redes sociales/web. Tesis de Maestría en Gestión de la Información. Valencia: Universitat Politècnica de València. Escola Tècnica Superior d’Enginyeria Informàtica. 2015. 50 p.

BARBERÁN, J.; RIVADENEIRA, F. y LARREA, J. Análisis de comentarios y revisiones de usuario usando redes neuronales recurrentes y procesamiento de lenguajes naturales. En: Revista de Tecnologías de la Informática y las Telecomunicaciones. Enero, 2018, vol.2, no.1, p.1-13

BARRIGA, José Camilo. Desarrollo y aplicación de una herramienta de extracción y almacenamiento de datos de Twitter a un contexto social de violencia política.

Trabajo de grado en Ingeniería de Sistemas. Bogotá, D.C.: Universidad Católica de Colombia, Facultad de Ingeniería. 2017. 106 p.

BERNERS, T.; CAILLIAU, R.; LUOTONEN, A.; NIELSEN, H. F. & SECRET. A. The world wide web. Communications of ACM, 37(8):76-82, 1994.

BIOLCHINI, J. et al., Systematic Review in Software Engineering. Systems Engineering and Computer Science Department, UFRJ: Rio de Janeiro, Brazil, 2005.

BOGARÍN, Alejandro; ROMERO, Cristóbal y CEREZO, Rebeca. Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. En: EDMETIC. 2016. vol.5, no.1, p.73-92

CAGNINA, Leticia; FERRETTI, Edgardo; VILLEGAS, M.Paula; GARCIARENA, M.José; BURDISSO, Sergio.; FÚNEZ, Darío; VELÁSQUEZ, Carlos y ERRECALDE, Marcelo. Miner´ıa de Textos y de la Web. En: XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). p. 268-272

CALVO, María. Text Analytics para Procesado Semántico. Tesis de Maestría en Técnicas Estadísticas. Santiago: Universidad de Vigo. 2017. 64 p.

CAMANA, Roberto. Potenciales aplicaciones de la minería de datos en Ecuador. En: Revista Tecnológica ESPOL – RTE, Julio 2016, vol. 29, no. 1, p.170-183.

CAMPAÑA, Francisco Xavier. Aplicación de técnicas de Data Mining a bases de datos de contenido musical para identificar rasgos de personalidad de los usuarios en el Distrito Metropolitano de Quito. Tesis de Magíster en Sistemas de Información e Inteligencia de Negocios. Sangolquí: Universidad de las Fuerzas Armadas, Vicerrectorado de investigación, innovación y transferencia de tecnología. 2017. 117 p.

CÁRCAMO, Luis; CALVA, Delia; RONQUILLO, Nayeli y NESBET, Felipe. México, en la prensa chilena: análisis basado en minería de datos textuales en Twitter. En: Revista Latina de Comunicación Social. 2017. no.72. p. 897 - 914.

CARRASCO-JIMÉNEZ, P. Análisis Masivo de Datos y Contraterrorismo. Tirant lo Blanch, Valencia, España, 2009.

CHAMBA, Sairy Fernanda. Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC. Trabajo de grado en Ingeniería de Sitemas. Loja, Ecuador: Universidad Nacional de Loja, Área de la Energía, las Industrias y los Recursos Naturales No Renovables. 2015. 184 p.

COELLO, Daniel. Análisis de conductas sociales aplicado a Big Data mediante técnicas de redes neuronales artificiales. Trabajo de grado en Ingeniería de

Sistemas Computacionales. Guayaquil, Ecuador: Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Físicas. 2017. 127 p.

COLLE, Raymond. Algoritmos, grandes datos e inteligencia en la red, una visión crítica. España: Univesidad de Alicante. 2017. 62 p.

COOLEY, R. MOBASHER, B. AND SRIVASTAVA. J. Data preparation for mining world wide web browsing patterns. Journal of Knowlegde and Information Systems, 1(1):5-32, 1999.

CONTRERAS, Dúber y MALDONADO, José. El papel de la minería de datos en la inteligencia de negocios, una revisión literaria. En: CIINATIC, Ponencia. 2017. p.1-8.

CONTRERAS, Marcial. Minería de texto en la clasificación de material bibliográfico. En: Biblios, 2016, no. 64. p. 33-43

CORCHERO, Pedro; FERNÁNDEZ, María y HURTADO, Ma. Antonia. Posicionamiento SEO del Centro de Investigación Flamenco Telethusa: Palabras claves. En: Revista Centro Investigación Flamenco Telethusa, 2016, vol.9, no.10, p.22-29.

CURIEL, Lorenzo y PANTOJA, Alayna. Estudio webmétrico de la revista electrónica Avanzada Científica. En: Revista de Arquitectura e Ingeniería, abril-2015, vol. 9, no. 1, p. 1-6

CURIEL, Silvio. Propuesta de indicadores y procedimientos para evaluar la usabilidad, comportamiento web y producción científica de las revistas del IDICT. En: Revista Documentación, 2014, vol. VII, no. 35, p.32-41.

DE FREITAS, Vidalina y YÁBER, Guillermo. Una propuesta de arquitectura para los Sistemas Informáticos de Gestión del Conocimiento en Instituciones de Educación Superior. En: Revista Espacios, 2015, vol.36, no.10, p.1-21.

DE LA CALLE, Guillermo. Modelo basado en técnicas de procesamiento de lenguaje natural para extraer y anotar información de publicaciones científicas. Tesis doctoral. Madrid: Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Informáticos, Departamento de Inteligencia Artificial. 2014. 191 p.

DURÁN, Zully y LEÓN, Kerly. Influencia del financiamiento en el desarrollo de las MYPES en el sector comercial del Distrito de Independencia - Período 2015. Trabajo de grado en Contaduría Pública. Huaraz, Perú: Universidad Nacional Santiago Antúnez de Mayolo, Facultad de Economía y Contabilidad. 2016. 122 p.

ECKERT, Karina; FAVRET, Fabián; BARBOZA, Matías; WITZKE, Leandro y ALVARENGA, Víctor. Modelos de Análisis de Información para la Toma de Decisiones Estratégicas del Sector Tealero. En: XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). p.117-121.

EIRINAKI, M.; and VAZIRGANNIS, M. Web mining for web personalization. ACM Transactions on Internet Technology, 3(1):1-27, February 2003.

ESPINOZA, Diego y ROJAS, Daniel. Minería de opiniones de usuarios en sitios web orientados al servicio. Trabajo de grado en Ingeniería civil informática. Valparaíso: Pontificia Universidad Católica de Valparaíso, Facultad de Ingeniería, Escuela de Ingeniería Informática. 2016. 189 p.

FALLOUX, Gonzalo. Diseño y desarrollo de un módulo de clasificación de páginas web en base a las características de su contenido utilizando técnicas de minería de datos. Trabajo de grado en Ingeniería Civil Industrial. Santiago de Chile: Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial. 2016. 120 p.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (ed.) Advances in knowledge and data mining. Cambridge (Massachussets): AAAI/MIT Press, 1996.

FERNÁNDEZ, Beatriz; DURÁN, Elena y AMANDI, Analía. Búsqueda y Recomendación de contenido educativo en entornos virtuales de aprendizaje. En: 15th Argentine Symposium on Articial Intelligence, ASAI 2014, p.67-74

GARCÍA, Alicia; FERRER, Antonia; PESET, Fernanda y GONZÁLEZ, Luis. Herramientas de análisis de datos bibliográficos y construcción de mapas de conocimiento: Bibexcel y Pajek. En: Bid, 2015, no. 34, p.1-7.

GARCÍA, Diego. Minería de datos aplicada a la enseñanza virtual: nuevas propuestas para la construcción de modelos y su integración en un entorno amigable para el usuario no experto. Tesis Doctoral. Cantabria: Universidad de Cantabria, Departamento de Ingeniería Informática y Electrónica. 2016. 294 p.

GARCÍA, D.F. Revisión sistemática de literatura en los trabajos de final de Máster y en las tesis Doctorales. Grupo de investigación en InterAcción y eLearning (GRIAL). España: Universidad de Salamanca. 2017.

GONZÁLEZ, Adela. Más allá del corpus: Big Data en la investigación lingüística: evolución, análisis y predicción del uso de la lengua a través de Twitter. Tesis Doctoral. Córdoba: Universidad de Córdoba, Departamento de Ciencias del Lenguaje, Área de Lingüística General. 2016. 392 p.

GONZÁLEZ, Carlos. Midiendo la calidad de la información gestionada: algunas reflexiones conceptuales-metodológicas. En: Biblios, 2014, no. 53, p.27-35.

GUANGA, Fernando y LONDOÑO, Julie. Interfaz gráfica para consulta de datos basada en operaciones de transformación de grafos. Trabajo de grado en Ingeniería de Sistemas y Computación. Cali, Valle: Pontificia Universidad Javeriana de Cali, Facultad de Ingenierías, Ingeniería de Sistemas y Computación. 2014. 170 p.

GUEVARA, Gladys; GUEVARA, Cristian y ELIZONDO, Daniel. Tratamiento de la información en la web: Text Mining y Web Mining. En: Revista Científica de Investigación actualización del mundo de las Ciencias, octubre, 2017, vol. 1, no. 4, p. 403-418.

GUTIÉRREZ, Jahir y MOLINA, Bernardo. Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. En: ONTARE: Revista de investigación de la Facultad de Ingeniería, 2015, vol.3, no.2, p.33-51.

GUZMÁN, Daniel. Minería de datos sobre opiniones de clientes en agencias virtuales de alojamiento. Trabajo de grado en Ingeniería Civil Informática. Valparaíso: Pontificia Universidad Católica de Valparaíso, Facultad de Ingeniería, Escuela de Ingeniería Informática. 2016. 130 p.

HASPERUÉ, Waldo. Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas. Tesis Doctoral en Ciencias Informáticas. La Plata: Universidad Nacional de La Plata, Facultad de Informática. 2012. 212 p.

HERNÁNDEZ, Emilcy; DUQUE, Néstor y MORENO, Julián. Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. En: TecnoLógicas, mayo - agosto, 2017, vol. 20, no. 39, p.1-24.

HIDALGO, Gustavo. Proyecto de detección de patrones de participación empleando minería de datos en el entorno virtual de aplicaciones web de la ESPOCH, para predecir estudiantes exitosos. Tesis de Magíster en Formulación, Evaluación y Gerencia de proyectos para el desarrollo. Riobamba, Ecuador: Escuela Superior Politécnica de Chimborazo. 2016. 140 p.

HOJAS, Wenny; SIMÓN, Alfredo y RODRÍGUEZ, Aramis. Aplicación de técnicas de minería de grafo para el análisis de textos. En: 18 Convención científica de Ingeniería y Arquitectura, 2017, p.1-10.

HURTADO, J. Metodología de la Investigación Holística. Caracas, Venezuela: Editorial SYPAL.

IZA, Miryan; SAQUICELA, Víctor y PACHECO, Idalia. Web Mining. En: Ciencia y Tecnología al servicio del pueblo, 2014, vol.1, no.3, p.134-139

JARAMILLO, Sonia; CARDONA, Sergio y FERNÁNDEZ, Alejandro. Minería de datos sobre streams de redes sociales, una herramienta al servicio de la bibliotecología. En: Inf. cult. soc. Ciudad Autónoma de Buenos Aires, dic. 2015, no.33, p.1-14.

JOHNSON, F. & KUMAR, S. Web Content Mining Techniques: A Survey. International Journal of Computer Applications (0975 – 888), June 2014, vol 47, no.11, p. 44-50.

LIM, Chee y CHINNASAMY, Balakumar. Conexión de los contenidos de la biblioteca utilizando minería de datos y análisis de texto en datos estructurados y no estructurados. En: IFLA WLIC, Mayo - 2013, p.1-14.

KITCHENHAM, B. Procedures for performing systematic reviews (Joint Technical Report). Software Engineering Group, Department of Computer Science, Keele University and Empirical Software Engineering National ICT Australia Ltd. 2004.

LOGICALIS. Modelos de data mining y las herramientas más usadas. Recuperado de Logicalis: Business and technology working as one, 2015: https://blog.es.logicalis.com/analytics/modelos-de-data-mining-y-las-herramientas-mas-usadas

MARKOV, Z. and LAROSE, D.T. Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage. John Wiley and Sons, New York, USA, 2007.

MARTÍNEZ, L.J. Cómo buscar y usar información científica: Guía para estudiantes universitarios. Santander, España: Universidad de Cantabria, 2013.

MENDOZA, M. Minería de datos en la web. Capítulo 19, 613-648, en CACHEDA, F.; FERNÁNDEZ, J.; y HUETE, J. Recuperación de información: Un enfoque práctico y multidisciplinar, Editorial Ra-Ma, 2011.

MOLINA, L.C. Torturando los Datos Hasta que Confiesen. Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña. Barcelona, España, 2000.

PINO, F.J.; GARCÍA, F. & PIATTINI, M. Revisión sistemática de mejora de procesos software en micro, pequeñas y medianas empresas. Revista Española de Innovación, Calidad e Ingeniería del Software, 2(1):6-23, 2006.

SCOTTO, M.; SILLITTI, A.; SUCCI, G.; VERNAZZA, T. “Managing Web-Based Information”, International Conference on Enterprise Information Systems (ICEIS 2004), Porto, Portugal, April 2004. Page 1-3

SOSA, M.O. & SOSA, E.C. Estudio de técnicas de Data Mining aplicadas al análisis de datos generados con la metodología Blended Learning. In XVI Workshop de Investigadores en Ciencias de la Computación. 2014.

SRIVASTAVA, J.; COOLEY, R.; DESHPANDE, M. and TAN, P.N. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, I(2):12-23, 2000.

TAVANI, H.T. Informational privacy, data mining, and the internet. Ethics and Information Technology, 1:137-145, 1999.

VEDDER, A. Privacy and con dentiality. medical data, new information technologies, and the need for normative principles other than privacy rules. Law and Medicine, 3:441{459, 2000.

VELÁSQUEZ, J.D. & DONOSO, L. Aplicación de Técnicas de Web Mining sobre los Datos Originados por Usuarios de Páginas Web. Visión Crítica desde las Garantías Fundamentales, especialmente la Libertad, la Privacidad y el Honor de las Personas. Revista Ingeniería de Sistemas, XIV: 47-68, Junio 2010.

VELÁSQUEZ, J.D. and PALADE, V. A knowledge base for the maintenance of knowledge extracted from web data. Knowledge-Based Systems, 20(3):238-248, 2007.

VILLATE, Jaime. Glosario de informática Inglés-Español. 2000. Recuperado de: http://quark.fe.up.pt/orca/pub-es/glosario.html.

ZABALA, F.R. Buscador de artículos científicos aplicando minería de datos. 2014. Instituto Tecnológico de la Paz.

ZAMORA, M.A. Internet. Universidad Autónoma del Estado de Hidalgo. 2014. P.3

ANEXO EJEMPLO MATRIZ RAE'S

#	BASE DOCUMENTAL	TÍTULO	AUTOR	PUBLICACIÓN	AÑO DE PUBLICACIÓN	TIPO DE DOCUMENTO
1	Redalyc	Revisión de lineamientos de accesibilidad y usabilidad para el diseño de sitios web para personas de la tercera edad	Aguirre, D.F. & Abadía, I.	Aguirre, D. & Abadía, I. (2017). Review of accessibility and usability guidelines for website design for the elderly people. Sistemas & Telemática, 15(42), 9-29.	2017	Artículo
2	Redalyc	La Web Oculta y cómo los buscadores encuentran la información	Amaro López, José Antonio; Lázaro, Ch. A.; Varela Navarro, Gerardo Alberto	Amaro López, JA, Lázaro, CA, Varela Navarro, GA. La Web Oculta y cómo los buscadores encuentran la información. Paakat: Revista de Tecnología y Sociedad [Internet]. 2014;(7).	2014	Artículo
3	Redalyc	Editorial: Data Mining in Electronic Commerce – Support vs. Confidence	Astudillo, César; Bardeen, Matthew; Cerpa, Narciso	Astudillo, C, Bardeen, M, Cerpa, N. Editorial: Data Mining in Electronic Commerce – Support vs . Confidence. Journal of Theoretical and Applied Electronic Commerce Research [Internet]. 2014;9(1):I-VII.	2014	Nota editorial
4	Redalyc	Minería de datos en egresados de la Universidad de Caldas	Bedoya, Oscar Mauricio; López Trujillo, Marcelo; Marulanda Echeverry, Carlos Eduardo	Bedoya, OM, López Trujillo, M, Marulanda Echeverry, CE. Minería de datos en egresados de la Universidad de Caldas. Revista Virtual Universidad Católica del Norte [Internet]. 2016;(49):110-124.	2016	Artículo
5	Google Scholar	Extracción de conocimientos de la Diabetes Tipo 1 utilizando la metodología de "Data Mining"	Abad, José María	Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación	2015	Tesis
6	Google Scholar	Búsqueda de patrones en el comportamiento de los visitantes de la plataforma "Oferto" de la Cámara de Comercio de Armenia y del Quindío, a través de la aplicación de minería web.	Angarita, D.A.; Muñoz, J.J.	Universidad Autónoma de Manizales	2016	Tesis
7	Google Scholar	Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística	Arcila-Calderón, Carlos; Barbosa-Caro, Eduar; Cabezuelo-Lorenzo, Francisco	Arcila-Calderón, Carlos; Barbosa-Caro, Eduar; Cabezuelo-Lorenzo, Francisco (2016). "Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística". El profesional de la información, v. 25, n. 4, pp. 623-631.	2016	Artículo
8	Google Scholar	Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación	Arias, A.; Mattos, Y.; Heredia, J.; Heredia, D.	A. Arias, Y. Mattos, J. Heredia & D. Heredia, "Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación", Investigación y Desarrollo en TIC, vol. 7, no. 1, pp. 14-20, 2016.	2016	Artículo
9	Google Scholar	Miner´ıa de Textos y de la Web	Leticia Cagnina, Edgardo Ferretti, M. Paula Villegas, M. Jos´e Garciarena, Sergio Burdisso,**Dar´ıo Funez, Carlos Vela´zquez, Marcelo Errecalde	Laboratorio de Investigaci´on y Desarrollo en Inteligencia Computacional Departamento de Inform´atica, Universidad Nacional de San Luis	2016	Artículo

10	Google Scholar	Aplicación de técnicas de minería de datos en redes sociales/web	Asensio Blasco, Elena	Universitat Politècnica de València	2015	Tesis
11	Google Scholar	Análisis de comentarios y revisiones de usuario usando redes neuronales recurrentes y procesamiento de lenguajes naturales.	Julio Barberan León; Fabricio Rivadeneira Zambrano; Jhonny Larrea Plua	Revista de tecnologías de la informática y las telecomunicaiones 2(1):1-13	2018	Artículo
12	Google Scholar	DESARROLLO Y APLICACIÓN DE UNA HERRAMIENTA DE EXTRACCIÓN Y ALMACENAMIENTO DE DATOS DE TWITTER A UN CONTEXTO SOCIAL DE VIOLENCIA POLÍTICA	JOSÉ CAMILO BARRIGA MARIÑO	UNIVERSIDAD CATÓLICA DE COLOMBIA	2017	Tesis
13	Google Scholar	Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle	Alejandro Bogarín Vega; Cristóbal Romero Morales; Rebeca Cerezo Menéndez	edmetic, 5(1), 2016, E-ISSN: 2254-0059; pp.73-92	2016	Artículo
14	Google Scholar	Text Analytics para Procesado Semántico	María Calvo Torres	Universidade de Vigo	2017	Tesis
15	Google Scholar	Aplicación de técnicas de Data Mining a bases de datos de contenido musical para identificar rasgos de personalidad de los usuarios en el Distrito Metropolitano de Quito.	Campaña Naranjo Francisco Xavier	Universidad de las Fuerzas Armadas	2017	Tesis
16	Google Scholar	México, en la prensa chilena: análisis basado en minería de datos textuales en Twitter	L Cárcamo-Ulloa, D Calva Rosales, N Ronquillo Rodríguez, , F Nesbet Montecinos	Revista Latina de Comunicación Social, 72, pp. 897 a 914.	2017	Artículo
17	Google Scholar	Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC	Sairy Fernanda Chamba Jiménez	Universidad Nacional de Loja	2015	Tesis
18	Google Scholar	ANÁLISIS DE CONDUCTAS SOCIALES APLICADO A BIG DATA MEDIANTE TÉCNICAS DE REDES NEURONALES ARTIFICIALES.	DANIEL DE JESÚS COELLO VARGAS	UNIVERSIDAD DE GUAYAQUIL	2017	Tesis
19	Google Scholar	Posicionamiento SEO del Centro de Investigación Flamenco Telethusa: Palabras claves	Pedro Corchero Murga; María del Rosario Fernández-Falero; Mª Antonia Hurtado Guapo	Revista Centro Investigación Flamenco Telethusa, 9(10), 22-29.	2016	Artículo
20	Google Scholar	Estudio webmétrico de la revista electrónica Avanzada Científica.	Curiel Lorenzo, Silvio; Pantoja Trincado, Alayna	Revista de Arquitectura e Ingeniería. 2015, Vol.9 No.1, 1-7	2015	Artículo

RESUMEN	OBSERVACIONES
Para el 2050 se estima en 10% el crecimiento de la población mayor en Colombia y con ello una mayor demanda de servicios especiales (como los de salud) para la tercera edad. Esto justifica que se exploren contenidos digitales de salud como una fuente importante de información para dicha población. Los lineamientos de accesibilidad y usabilidad para el diseño de sitios web –e.g., TAW y WACG– no poseen lineamientos específicos para mitigar las discapacidades motrices, cognitivas o visuales, propias del envejecimiento, que se convierten en una barrera para que este grupo consulte información necesaria para los procesos administrativos que involucran su salud. Se presenta esta revisión de lineamientos de accesibilidad y usabilidad, que facilitan el consumo de contenidos específicos y generan mejores interacciones con dichos sistemas, lo que propiciará la construcción de lineamientos basados en recomendaciones ya existentes que permitan desarrollar aspectos relacionados con la interacción, legibilidad y usabilidad en contenidos digitales para personas de la tercera edad.	Web Content Accessibility Guidelines [WCAG] La W3C, basada en los siete principios de diseño Universal, propone cuatro pautas en las que se categorizan los niveles de accesibilidad y usabilidad, los lineamientos de la iniciativa WCAG establecen los protocolos para diseño inclusivo: perceptible, operable, comprensible y robusto. Estas pautas contienen, a su vez, criterios de conformidad: A, pauta que atribuye mayor importancia en cuanto a accesibilidad final; AA, elimina importantes barreras de acceso a la web; y AAA, menor importancia, pero confiere buen nivel de accesibilidad (Oleo & Rodríguez, 2013). Las normas de accesibilidad web con mayor aceptación son las de la WCAG 1.0, publicadas en 1999 por la W3C (actualizadas en 2008), ellas adaptan los contenidos a las nuevas tecnologías y mejoran la aplicación e implementación de los lineamientos y criterios de accesibilidad y usabilidad propuestos por la versión anterior (W3C, 2008; Oleo & Rodríguez, 2013).
En el presente trabajo se aborda la diferencia entre lo que es la Web de la superficie y la Web Oculta, así como los problemas que en la actualidad deben resolver los buscadores para lograr indexar en sus bases de datos la mayor cantidad posible de sitios de la Internet, para proveer a sus usuarios de páginas que satisfagan sus necesidades de información.	Para hablar de la Deep Web primero debemos abordar lo que es la web de la superficie o Surface Web, la cual, según (López-Barberá Martín , 2014, p. 96), es el conjunto de páginas que en la actualidad podemos consultar mediante los buscadores; es decir, todas aquellas páginas que no se encuentran elaboradoras en HTML1, CSS2, y que no contengan un formulario para acceder al contenido, además de que puedan ser indexadas por los buscadores mediante métodos que hacen uso del seguimiento de los enlaces que estas páginas contienen. Ahora bien, la Deep web también llamada web oculta -la internet oculta- es la parte de internet a cuya información no es posible acceder de manera total mediante los buscadores, porque no es posible indexar las páginas de los sitios. Lo anterior debido a que el acceso, a las mismas, se encuentra restringido, ya sea por contraseña -como ocurre con los correos electrónico o sistema de bases de datos en línea de las empresas o de instituciones de gobierno- o mediante el llenado de un formulario que le permite al usuario solicita información para acceder a ésta; sirva de ejemplo, para este último caso, las bibliotecas virtuales y los formularios de páginas de comercio electrónico. La información pública en la Deep Web es de 400 a 500 veces más extensa que la contenida en la web de la superficie. La Deep Web contiene 7,500 terabytes de información en comparación con la web de la superficie que sólo contiene 99 terabytes. La Deep Web contiene cerca de 550 billones de documentos individuales comparado contra un billón que se localizan en la web de la superficie. Existían en el año 2000 más de 200,000 sitios web ocultos. Sesenta de los sitios más grandes de la Deep Web contenían cerca de 750 terabytes de información, lo que excedía 40 veces el tamaño de la información contenida por todos los sitios en la web de la superficie. En promedio, los sitios en la Deep Web reciben arriba del 50% del tráfico mensual que los sitios contenidos en la web de la superficie.
	El proceso de minería de datos implica buscar, seleccionar, explorar y modelar grandes cantidades de datos para descubrir patrones previamente desconocidos que son potencialmente útiles y, en última instancia, información comprensible, desde grandes bases de datos. Su objetivo es manipular los datos en conocimiento. La extracción de patrones es una importante proceso de cualquier técnica de minería de datos y se refiere a las relaciones entre subconjuntos de datos. La minería de datos utiliza diferentes familias de métodos computacionales, estadísticos y de aprendizaje automático que incluyen estadísticas análisis, árboles de decisión, redes neuronales, inducción y refinamiento de reglas, y visualización gráfica entre otros, para explore exhaustivamente los datos para revelar relaciones complejas que puedan existir. Aunque las técnicas de aprendizaje automático tienen estado disponible desde hace mucho tiempo, el desarrollo de herramientas avanzadas y fáciles de usar para la inteligencia empresarial ha hecho que la minería de datos sea más atractiva y práctica para las organizaciones. Cuando estas técnicas de extracción de patrones son usados correctamente, pueden ser herramientas efectivas para extraer información útil de los datos.
Este artículo presenta los resultados del uso de técnicas de clasificación en minería de datos de los factores asociados a la percepción que el recién egresado de la Universidad de Caldas tiene de la utilidad de los conocimientos y destrezas adquiridos a lo largo de sus estudios, que forman parte vital en su rol laboral. Para su desarrollo se utilizaron enfoques investigativos como el exploratorio y el descriptivo, conjuntamente con cuatro técnicas de minería de datos de clasificación y un repositorio de datos con información del entorno social, personal y familiar, académico, laboral y de percepción de utilidad de habilidades frente al reto profesional. El rasgo del egresado determina que las habilidades y destrezas adquiridas durante sus estudios en la Universidad de Caldas son muy útiles; se espera que constituyan un aporte significativo en la búsqueda de mecanismos que mejoren el nivel de satisfacción de los egresados con su formación y la pertinencia de los planes de estudio.	Según Pérez-Palacios, Caballero, Caro, Rodríguez y Antequera (2014), la minería de datos es una parte importante de un proceso más amplio conocido como descubrimiento de conocimiento en bases de datos (KDD por sus iniciales en inglés). El objetivo principal de la minería de datos consiste en extraer información oculta de un conjunto de datos. Esto puede ser alcanzado por el análisis automático o semiautomático de gran cantidad de datos, lo que permite la extracción de patrones desconocidos. Estos patrones pueden ser grupos de registros de datos (análisis clúster), inusuales registros (detección de anomalías) y dependencias entre datos (asociación de reglas). Por lo tanto, los patrones pueden ser vistos como un resumen de los datos de entrada, y se pueden utilizar para su posterior análisis. Tsai (2013) complementa lo anterior en tanto explica que la minería de datos es un campo interdisciplinario que combina la inteligencia artificial, la gestión de bases de datos, visualización de datos, aprendizaje automático, algoritmos matemáticos y estadísticos. Esta tecnología ofrece diferentes metodologías para la toma de decisiones, resolución de problemas, el análisis, la planificación, el diagnóstico, la detección, la integración, la prevención, el aprendizaje y la innovación. En esta misma línea, Natek y Zwilling (2014) explican los pasos para el análisis de la MD: primero, crear los datos conjuntos; segundo, definir la herramienta de MD a utilizar; tercero, evaluar las técnicas de MD a utilizar; y cuarto, analizar los datos por cada modelo y elegir el mejor. En la minería de datos los procedimientos de clasificación desarrollan un modelo agregado por reglas (si-entonces) y se aplican cabalmente. En concordancia, el efecto de aplicar el algoritmo de clasificación se direcciona a comparar la clase predicha con la clase real de las instancias. Este proceso de minería de datos busca reglas para definir si un ítem o un evento pertenecen a una clase de datos en particular. Para el conjunto de datos se cuenta con un conjunto apropiado de atributos predictivos, de tal manera que el modelo busca identificar los egresados con mayor propensión a justificar una percepción "muy útil", "poco útil" o "nada útil" de las destrezas adquiridas en algún programa cursado.
Data mining, también referenciado como descubrimiento de consnocimiento en bases de datos (Knowledge Discovery in Database o KDD), ha sido definida como el rpoceso de extracción no trival de información implícita, previamente desconocida y potencialmente útil. Todo ello, sirviéndonos de las siguietnes fases de extracción del conocimiento: selección de datos, pre-procesado, transformación, minería de datos, interpretación de los resultados, evaluación y obtención del conocimiento. Con todo este proceso buscamos generar un único modelo insulina glucosa que se ajuste de forma individual a cada paciente y sea capaz, al mimso timepo, de precedir los estados futuros glucosa con cálculos en tiempo real, a través de unos parámetros introducidos.	Técnicas de minería de datos: La regresión lineal, métodos basados en núcleo, métodos bayesianos, árboles de decisión, técnicas de conteo y soporte mínimo, redes neuronales artificiales, aprendizaje basado en instancias o casos, algoritmos evolutivos
En la actualidad las organizaciones cuentan con información generada cada vez más rápido y de manera exponencial por el uso de las páginas web por parte de los usuarios. Dicha generación de información se debe a la publicación de sus productos y/o servicios en sus sitios web y la interacción de los usuarios con los mismos, es por esto que se wuelve necesario el análisis de dicha información con el objetivo de ser competitivo y obtener utilidades, usando como medio el mundo digital. Tal como afirma Baeza-Yates (2005) "la información de la web es finita pero el número de páginas web es infinito", a partir de esta premisa es claro que se cuenta con información valiosa para la gestión de la Organización. Sin embargo, para que esta información pueda tener un impacto adecuado se debe realizar un proceso con las técnicas apropiadas, ya que la mayoría de veces la información importante no se encuentra a simple vista y si no es utilizada y explotada de la forma correcta o simplemente no se hacen las búsquedas adecuadas dentro de la misma, termina por convertirse en datos sin valor.	Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. Esta tesis fue presentada por Juan Miguel Moine de la Universidad Nacional de La Plata en Buenos Aires, Argentina (2013). El objetivo de este trabajo fue realizar un análisis comparativo entre las metodologías de minería más difundidas. Para realizar dicho análisis se elaboró un cuadro comparativo con las características de cada metodología. Con el cuadro en mención se analizó el nivel de especificación de las tareas, los escenarios de aplicación, las actividades que componen cada fase del proceso y las actividades destinadas a la dirección del proyecto. Metodología Cross-Industry Standard Process for Data Mining (CRISP DM): El desarrollo del proyecto planteado estará basado en la aplicación de la metodología Cross-Industry Standard Process for Data Mining (CRISP DM) probada para orientar trabajos de minería de datos. Si bien el Data Mining y la Web Mining tiene similitudes, existen grandes diferencias que las separa a la hora de realizar el análisis de los datos. Esto se debe a que la Web es poco estructurada por naturaleza a diferencia de las bases de datos relaciones. Esto provoca que las técnicas de la minería de datos no se puedan aplicar directamente sino que deban modificarse para superar el problema de estructura. Algunos otros problemas que presentan los datos en la Web son los siguientes: La falta de contexto en la información y en las bases de datos. Separar la información relevante de que no lo es. Sobrecarga de información. Es por esta razón que, según Salton, G. y McGill, M. J., (1983), la minería Web toma elementos como la recuperación de información teniendo en cuenta el procesamiento del lenguaje natural (Natural Language Processing - NLP) la inteligencia artificial y el aprendizaje automático, que son propios de otras áreas de investigación. Descripción amplia de web mining.
Este trabajo conceptualiza el término big data y describe su importancia en el campo de la investigación científica en ciencias sociales y en las prácticas periodísticas. Se explican técnicas de análisis de datos textuales a gran escala como el análisis automatizado de contenidos, la minería de datos (data mining), el aprendizaje automatizado (machine learning), el modelamiento de temas (topic modeling) y el análisis de sentimientos (sentiment analysis), que pueden servir para la generación de conocimiento en ciencias sociales y de noticias en periodismo. Se expone cuál es la infraestructura necesaria para el análisis de big data a través del despliegue de centros de cómputo distribuido y se valora el uso de las principales herramientas para la obtención de información a través de software comerciales y de paquetes de programación como Python o R.	Implica la extracción de conocimiento a partir de datos masivos y las relaciones subyacentes que pueden existir entre ellos. El data mining se originó en 1990 a medida que la tecnología relacional de bases de datos maduró y los procesos de negocio crecieron en automatización (Dhar, 2013, p. 67), fomentando la creación de software orientado a aprovechar los datos sobre comportamiento y transacciones, para predecir y planear de manera más acertada. Siguiendo la línea de Han, Kamber y Pei (2006), el knowledge discovery from data (término que se ha usado a la par de data mining) se puede dividir en siete fases: - limpieza de datos - integración de los datos - selección de datos - transformación de los datos - minería de datos - evaluación de patrones - presentación del conocimiento.
Por medio de la extracción de datos tanto precisos como no relevantes sobre un texto o estructura de datos textual, se es posible generar nueva información y conocimiento, este análisis es realizado con la minería de texto. Si bien esta práctica puede conducir más apropiadamente hacia la generación y/o descubrimiento de nuevos datos, también es viable para la verificación y elaboración de datos a referenciar. Ya que su uso yace en el análisis de la información arrojando datos de manera indiscriminada, esto resulta útil a la hora de rectificar la compatibilidad de la información, entre dos o más medios de contenido textual. Se expondrán y explicarán las diversas maneras como a través de la minería de texto, ayuda en la elaboración de un artículo de investigación o un estado del arte, teniendo en cuenta el tópico a explicar en dicho texto, se utilizarán diferentes metodologías para implementar la minería de texto con el fin de analizar otros diferentes artículos potenciales referencias, teniendo así, un mejor enfoque sobre el material cimiento de un nuevo proyecto o artículo. Adicional a esto se mostraran pruebas realizadas en la herramienta de software Weka donde se evidencian los resultados que arroja el software de los artículos que se analizan por medio de este.	Lo primero que se debe tener en cuenta la función que cumple la minería de texto es la recuperación de información, es decir, con la extracción le dicha información se hace la construcción de una base de datos la cual permitirá el cotejo de estas mismas. Esto puede ser aplicado en diferentes campos en donde se encuentre gran cantidad de texto, desde mensajes de Twitter a una colección de artículos científicos, dependiendo del tema que sea de gran interés. Luego de esta labor de recopilación, se reconocen entidades nombradas. Esto hace referencia a identificar las partes especificas del texto y todas aquellas pistas que me permitan generar información para llegar al objetivo y descartar aquellas palabras que tiene doble significad diferente a lo que se quiere. La minería de texto es una de las mejores herramientas que permite la extracción de dicha información en un texto de interés de esta manera a través de una aplicación sistematizada se puede lograr obtener información en textos de forma rápida y de manera eficiente se pueden construir conceptos a partir de esta búsqueda minuciosa.
Este art´ıculo describe, brevemente, las tareas de investigaci´on y desarrollo que se est´an llevando a cabo en la l´ınea de investigaci´on "Miner´ıa de Textos y de la Web" en el marco del proyecto "Aprendizaje autom´atico y toma de decisiones en sistemas inteligentes para la Web". La l´ınea aborda diversas ´areas vinculadas a la ingenier´ıa del lenguaje natural, como por ejemplo el Procesamiento del Lenguaje Natural (PLN), la Ling¨u´ıstica Computacional, la Miner´ıa de Textos, la Miner´ıa de la Web y la recuperaci´on de informaci´on de la Web. En el contexto de este proyecto por lo tanto, esta l´ınea se centra en todos los problemas vinculados con el desarrollo de herramientas inteligentes para la extracci´on, an´alisis y validaci´on de contenido Web, que incluyen: representaci´on de documentos y usuarios de la Web, medidas de calidad de informaci´on para el contenido Web, t´ecnicas abiertas de extracci´on de informaci´on para la Web, algoritmos de categorizaci´on supervisados, semi-supervisados y no supervisados y caracterizaci´on de usuarios, entre otros.	Debido al f´acil acceso a la informaci´on que existe en la actualidad a trav´es de diferentes recursos, la evaluaci´on de la calidad de la informaci´on en la Web se ha convertido en una tarea muy importante. D´ıa a d´ıa tanto las personas comunes como empresas y entidades gubernamentales o privadas toman decisiones bas´andose en la informaci´on disponible en la Web. Esto, sumado al notable incremento de informaci´on disponible en Internet ha provocado una necesidad imperiosa de evaluar la calidad de dicha informaci´on de forma autom´atica.

Las técnicas de minería de datos permiten obtener información de redes sociales como Twitter. Su análisis correcto proporciona un valor adicional a la recuperación de información. El procesado de lenguaje natural, así como la delimitación geográfica del origen de sus mensajes, se convierte en un objetivo fundamental y punto de partida de cualquier investigación. Por ello, nuestro estudio se basa en la realización de una aplicación que gestione y obtenga datos sobre el uso de los lenguajes oficiales de las comunidades autonómicas en Twitter	El análisis lingüístico de los textos de Twitter de forma automática se ha basado en la librería NLTK. Pero en las stopwords que proporciona este módulo no figuraban las lenguas autonómicas, por tanto, la búsqueda de estas palabras vacías en otros idiomas ha supuesto revisar alternativas para su obtención, y se han tenido problemas con la codificación de estos archivos y el estándar del módulo. En cuanto a proporcionar una visualización significativa de los datos obtenidos, si bien el mapa cumple con sus objetivos de marcar la localización de tweets, seguir un paso más allá y hacer la web más interactiva, hubiera facilitado de cara al usuario una mayor comprensión de la aplicación y su reutilización de formas diversas.
La minería de opinión tiene como principal herramienta el procesamiento de lenguajes naturales y mediante técnicas linguisticas computacionales realizar grandes series de operaciones sobre masivas cantidades de texto para un objetivo común. Esta investigación se centra en la minería de opinión y la utilización del procesamiento de lenguajes naturales y redes neuronales para identificar y extraer información subjetiva de los entornos que se presenten para su análisis, bajo un grado de polaridad comprobar la connotación que esta técnica puede influir en toma de decisiones económicas que puedan estar relacionada con la información. Se hará uso de la red social Facebook, la red social más grande de la cual tiene un promedio de 10.2 millones de comentarios cada 20 minutos y mediante Amazon una de las compañías mas grandes de comercio electrónico del mundo la cual publica contenido comercial alojado en Facebook poder analizar el contenido sintáctico bajo un nivel connotativo de polaridad en correspondencia con la información relacionada en Amazon estableciendo categorías de productos que se ofrecen en Facebook y sus influencias en las revisiones que tienen los mismos productos ofrecidos en Amazon y determinar el grado de relación sobre las preferencias que tienen los usuarios bajo estas tecnologías.	La minería Web se la utilizo para la recuperación de información mediante la extracción de un conjunto de documentos XML que son generados para cada usuarios de la red. Previamente la minería web se basa en la clasificación del conjunto de documentos por lo general de acuerdo con sus puntuaciones de relevancia.[1] En la siguiente figura se muestra el proceso de captura y transformación de este tipo de información.
Este proyecto se orientó a la construcción de una herramienta web para la extracción y almacenamiento de datos de la red social twitter, la cual permitirá a futuro con apoyo de un software externo o integrado, establecer un análisis estadístico de estos datos, enfocado en la necesidad del usuario. En este caso, las funcionalidades de la herramienta se aplican a un contexto social enfocado en la medición de violencia política en twitter. Para su realización se integraron bases teóricas enfocadas en conceptos relacionados con BigData, Almacenamiento y minería de datos, haciendo énfasis en los procesos de extracción, almacenamiento y curación de datos. Así mismo, se realizó un proceso de desarrollo de software en el cual se aplicó la metodología PXP, la cual es una adaptación de la metodología de programación extrema enfocada al desarrollo llevado a cabo por un solo programador. La cual permitió de forma ágil implementar la herramienta y generar como resultado, un set de pruebas unitarias con las cuales se identifica el correcto funcionamiento del software. La conectividad de la herramienta con la red social especificada implicó el uso de tecnologías como interfaces de programación y servicios web, las cuales permitieron que los procesos ligados a la extracción de datos o minería web se realizaran de forma correcta y así generar un contexto de información abierta al análisis. Información que sin importar las limitaciones con las cuales fue desarrollada la herramienta, supera en valor de utilidad y variedad a la información que ofrecen diferentes herramientas del mercado. El sistema extrae y almacena los datos de las cuentas requeridas las cuales son previamente ingresadas al sistema y permite al usuario gestionar una clasificación enfocada al contexto social tratado, posteriormente esta información provee un ambiente de calidad para el análisis estadístico requerido y que abre las puertas a la construcción de diferentes soluciones.	La minería de contenido web es el proceso de extraer información útil del contenido de los documentos web. Puede consistir en texto, imágenes, audio, vídeo, o registros estructurados, tales como listas y tablas. La aplicación de la minería de texto de contenido web ha sido el más ampliamente investigado. Los temas abordados en la minería de texto incluyen descubrimiento y seguimiento de temas, la extracción de patrones de asociación, agrupación de documentos web y clasificación de páginas web. Investigaciones sobre este tema se han basado en gran medida en las técnicas desarrolladas en otras disciplinas tales como la Recuperación de Información (RI) y Procesamiento del Lenguaje Natural (PLN).(Srivastava, Desikan, & Kumar, 2006)
En este artículo, aplicamos técnicas de minería de datos para descubrir rutas de aprendizaje frecuentes. Hemos utilizado datos de 84 estudiantes universitarios, seguidos en un curso online usando Moodle 2.0. Proponemos agrupar a los estudiantes, en primer lugar, a partir de los datos de una síntesis de uso de Moodle y/o las calificaciones finales de los alumnos en un curso. Luego, usamos los datos de los logs de Moodle sobre cada cluster/grupo de estudiantes separadamente con el fin de poder obtener más específicos y precisos modelos de procesos del comportamiento de los estudiantes.	Process Mining (PM) (Trcka y Pechenizkiy, 2009) es una técnica para hacer minería de datos sobre las aplicaciones que generan registro de eventos para identificar posibles procesos en una variedad de dominios de aplicación. La aplicación de las actividades de la minería de procesos debe tener como resultado modelos de flujos de procesos de negocio y de información de su empleo histórico (caminos más frecuentes, actividades menos realizadas, etc.). Herramientas de PM como ProM (VAN DER AALST, 2011) brindan análisis y descubrimiento de flujos de procesos a partir de los registros de eventos generados por muchas aplicaciones.
Este trabajo trata de introducir las principales técnicas de la minería de textos y sus diversas aplicaciones. Para ello, en el primer capítulo, se comienza a explicar qué es la minería de textos y cuáles son sus principales usos actualmente. Seguidamente, se explican varias formas de cómo obtener los datos y luego se ilustra el preprocesado de los mismos con un ejemplo. En el tercer capítulo, se hace un análisis exploratorio de los datos y, a continuación, se explican los métodos más utilizados para la clasificación de documentos, se entrenan los modelos y se validan. Por último, se muestra una aplicación Shiny del clasificador de textos que se ha construido siguiendo paso a paso las explicaciones del trabajo. Su finalidad es una mejor comprensión del tema con la ayuda de un método visual interactivo. Para la realización de este trabajo, se ha recibido la ayuda de la empresa Oesía, experta en el ámbito de la minería de textos.	Existen dos tipos de clasificación, la supervisada y la no supervisada. La primera es en la que la respuesta es conocida y en la segunda, también llamada cluster, no se tienen las posibles respuestas. Cada una de ellas tiene sus ventajas e inconvenientes, por lo que dependiendo de cuál sea el objetivo, se emplea una u otra. La clasificación supervisada proporciona a los investigadores la oportunidad de especificar las categorías para el algoritmo de clasificación. El inconveniente es que estos clasificadores requieren mano de obra, debido a las grandes cantidades de datos de entrenamiento necesarias. El concepto de datos de entrenamiento se explicará en el próximo apartado. Los métodos de clasificación no supervisada generan ellos mismos categorías. Esto es interesante para el estudio de las principales líneas de división en un corpus lingüístico. Además, no necesita una codificación manual de los datos. Dos de los contras principales de esta clasificación son que se requiere que los investigadores especifiquen el número de categorías en las que se va a agrupar el corpus o que establezcan un criterio que determine ese número, y, además, la interpretación a posteriori de los resultados de la estimación suele ser muy compleja.
El presente proyecto busca encontrar dichos datos arraigados en bases de datos de contenido musical y correlacionarlos con los rasgos de la personalidad de los individuos haciendo uso de técnicas y algoritmos de minería de datos para generar información que pueda ser de utilidad en distintos campos de acción como son la psicología, recursos humanos, marketing, etc. Posteriormente es necesario identificar un mecanismo para poder evaluar los rasgos de personalidad. Esto se enmarca dentro de cinco grandes factores ya determinados por la psicología a lo largo de años de estudios previos centrados en el modelo OCEAN.	La minería de datos o exploración de datos (es la etapa de análisis de "Knowledge Discovery in Databases" o KDD) es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y de gestión de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de Intereses, de consideraciones de la Teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea. La minería de datos es la aplicación de técnicas de minería de datos para descubrir los patrones de la Web. De acuerdo a los objetivos de análisis, la minería web se puede dividir en tres tipos diferentes, que son la minería de uso de la Web, minería del contenido de la Web y minería de la estructura de la Web. Minería del contenido de la Web es el proceso de descubrir información útil de texto, imagen, audio o datos de vídeo en la web. La minería de contenido web a veces se llama la minería de textos web, porque el contenido del texto es la zona más ampliamente investigado. Las tecnologías que se utilizan normalmente en la minería de contenido web son PNL (procesamiento de lenguaje natural) e IR (recuperación de información). Detalle de minería web
El presente artículo investiga 3.842 menciones referidas a México y los mexicanos en la prensa chilena (290 medios), basándose en sus titulares en un período comprendido entre el 15 marzo y el 15 de junio de 2016. La colección de datos se extrajo desde las cuentas de los medios en la red social Twitter. El objetivo general del estudio es describir cuánto se habla y sobre qué temas en la prensa chilena sobre la nación mexicana, así como el impacto que esto puede generar en la construcción de la imagen intercultural. Para este trabajo se utilizó una metodología basada en minería de datos, a partir de un crawler (software automatizado que recolecta metódicamente datos textuales y frecuencias de emisión) que sigue a los medios de prensa chilenos y almacena sus titulares en un servidor. Posteriormente, con el apoyo de herramientas tales como Elasticsearch y Kibana, se pudo explorar los datos textuales contenidos en la base de datos. Con posterioridad al minado clasificamos los titulares en 10 frentes informativos. Entre las conclusiones podemos señalar que existe una fuerte asociación de México a los acontecimientos deportivos, a los problemas de narcotráfico, violencia y sucesos protagonizados por artistas.	Normalmente toda la actividad de un usuario en la red deja registro en las bases de datos de los servidores web. En las redes sociales, los medios de comunicación participan desde cuentas con similares condiciones técnicas y construyen públicos destinatarios más reducidos que en sus versiones tradicionales de papel o televisión, pero igual de ávidos por consumir información. En este contexto, la ciencia de datos web o minería de datos aparece para extraer y comprender toda aquella información útil, producida por los usuarios. Las herramientas de minería de datos analizan esta ingente información y la convierten en conocimiento. "La minería de datos es, esencialmente, un proceso conducido por un problema: la respuesta a una pregunta o la solución a un problema se busca analizando los datos disponibles. El análisis de los datos forma el núcleo de la minería de datos, pero el proceso completo abarca también temas tales como la definición del problema empresarial y el desarrollo de la solución para resolverlo" (González, 2007, p. 1). En este sentido, el objetivo de la minería de datos es aprovechar la hiperabundancia de información. Por eso para el trabajo en esta técnica es muy relevante establecer términos de búsqueda concretos, puesto que en un océano de información es fácil desviar el sentido y perder de vista el objetivo inicial. Estrategias de análisis Dado que la extracción desde Twitter arrojó un corpus manejable que podía ser analizado manualmente se procedió a realizar dos tipos de análisis Análisis estadístico-descriptivo asistido por tecnologías web. A través de visualizaciones de datos obtenidas gracias a las herramientas: ElasticSearch y Kibana (aplicaciones asociadas al crawler del proyecto) se obtuvieron a) histogramas que dan cuenta de la evolución temporal de los vocablos en el corpus de análisis y b) gráficos de composición que permiten identificar cuáles son los medios y en qué porcentaje usan las palabras buscadas. Análisis de contenidos mediado por arbitraje de jueces. Dos jueces expertos, periodistas y estudiantes de magíster en comunicación clasificaron los titulares en 10 categorías temáticas: accidentes, deportes, ecología, economía, entretenimiento, judicial, política, salud, sociedad y tecnología. siguiendo como modelo la clasificación establecida por Vernier, Cárcamo y Scheihing, (2016). Además, al interior de cada categoría se identificaron temas recurrentes en cada subgrupo. Los resultados de estos arbitrajes también se contabilizaron y expresaron en gráficos de composición
La Minería de Datos aplicada en el ámbito de la comercialización permite entre otros aspectos descubrir patrones de comportamiento de clientes, que las empresas pueden utilizar para elaborar estrategias de marketing dirigidas hacia los distintos tipos de clientes que poseen. El agrupamiento o clustering representa una de las técnicas de Minería de Datos más utilizada para este tipo de análisis, esta técnica se basa en la división de un conjunto de datos en pequeños segmentos o grupos, en donde cada segmento contiene datos similares dentro de sí y mantiene una marcada diferencia con respecto a los otros segmentos. El presente Trabajo de Titulación tiene por objetivo obtener la segmentación de clientes en la empresa tecnológica Master PC mediante la aplicación de técnicas de Minería de Datos, para ello se tomó en cuenta el comportamiento de compra, que permitió identificar la lealtad de los clientes de la empresa tecnológica Master PC. Se aplicó la metodología CRISP-DM para el proceso de Minería de Datos. El análisis se realizó en base al modelo RFM (Recencia, Frecuencia, Valor Monetario), y sobre este modelo se aplicaron los algoritmos de agrupamiento: k-means, k-medoids, y Self-Organizing Maps (SOM). Para validar el resultado de los algoritmos de agrupamiento y seleccionar el que proporcione grupos de mejor calidad, se ha aplicado la técnica de evaluación en cascada aplicando un algoritmo de clasificación. Finalmente se utilizó el algoritmo Apriori para encontrar asociaciones entre productos, para cada grupo de clientes. La herramienta utilizada para el proceso de Minería de Datos fue el entorno RStudio.	Las técnicas de Minería de Datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos. Las principales técnicas de Minería de Datos se suelen clasificar según su tarea de descubrimiento [10]. De acuerdo a esto las técnicas de Minería de Datos se clasifican en dos grandes categorías: predictivas y descriptivas [2]. Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de Minería de Datos, antes de aceptarlo como válido [6]. Se trata de problemas y tareas en los que hay que predecir uno o más valores para uno o más ejemplos [11]. Los ejemplos en la evidencia van acompañados de una salida (clase, categoría o valor numérico) o un orden entre ellos. Dependiendo de la correspondencia entre los ejemplos y los valores de salida y la presentación de los ejemplos, podemos definir varias tareas predictivas que se describen a continuación. Detalle de las técnicas
Las distintas empresas que manejan un foro donde interactuan diferentes usuarios, sin importar la étnia, localización o lenguaje nativo, realizan diferentes conjuntos de procesos para efectuar una buena comunicación entre usuarios a la hora de compartir ideas, recursos u opiniones. Algunas empresas – instituciones, regulan estos sitios web mediante tecnicas tecnológicas tradicionales, por ende, el tiempo de análisis podría no ser el más optimo. El realizar un análisis y clasificar los comentarios de cientos, miles e inclusive millones de usuarios es una tarea demandante, tedioso y cansado. Para esto, se desarrolla un algoritmo encargado de clasificar y mostrar gráficamente los resultados del análisis dentro de la base de datos, este estudio sigue los principios de una de las ramas de la Inteligencia Artificial "Las Redes Neuronales Artificiales"…	Es la razón por la cual se define a la minería de datos como: "Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" Según (Fayad, 1996, pág. 187). A su vez se afirma que: "Numerosos especialistas señalan que la Minería de Datos necesita de la integración de enfoques de múltiples disciplinas" (Mitchell, 2006. pág. 120). También podemos notar que "Gracias a la computación ha surgido la minería de datos, la cual consiste de algoritmos que extraen conocimiento de grandes bases de datos que acumulan la historia de las actividades de las organizaciones." Zapata, G. A. V. El conocimiento tiene como finalidad prevenir a los responsables de tomar decisiones sobre situaciones interesantes, anomalías e incluso amenazas no detectadas con anticipación. Los llamados "mineros son auxiliares indispensables para la dirección de cualquier organización".
El estudio, clasificación y organización de los metadatos extraídos de los sitios web es una herramienta de gran utilidad para optimizar su posicionamiento. El objetivo de este trabajo es posicionar el web site del Centro de Investigación Flamenco Telethusa. Para ello se han analizado los términos utilizados en los buscadores para encontrar el sitio web de la entidad. Los 10 términos más usados para la búsqueda según el sistema de estadísticas AWStats son los siguientes: flamenco, baile, danza, ballet, estructura, lesiones, telethusa, alegrías, bailar, pie. El uso de estos términos de búsqueda ha sido analizado y compilado por meses y años, desde diciembre de 2008 hasta diciembre de 2014. Los resultados de búsqueda para cada uno de estos términos han sido: flamenco, 9600; baile, 4196; danza, 3147; ballet, 1819; estructura, 1651; lesiones, 989; telethusa, 888; alegrías, 887; bailar, 803; y pie, 851. El número de búsquedas del sitio ha ido aumentando de forma progresiva desde 2009 hasta año 2012 donde alcanza su máximo. Se produce una disminución durante los años 2013 y 2014. El mes de noviembre, para todos los años, es en el que se produce un mayor número de búsquedas, y el término flamenco el más buscado. Como conclusión proponemos que sea en este mes de noviembre cuando se publiquen contenidos nuevos para satisfacer la demanda de los usuarios.	La minería Web permite recuperar información de un sitio web de manera automática (Fuentes Reyes, 2007), hecho de gran valor para las organizaciones y muy utilizado para el estudio del posicionamiento Web y el comercio electrónico. La Minería Web de contenido es la que clasifica y organiza los metadatos extraídos, con el objeto de recuperar y facilitar el acceso a la información. Tales metadatos pueden ser estudiados estáticamente mediante instantáneas de la web en un periodo determinado. La minería de datos consta de cuatro fases: recolección automática de la información importante para procesarla posteriormente; procesamiento de datos, ordenándolos y clasificándolos automáticamente; descubrimiento de patrones mediante hallazgo de frecuencias, reglas de asociación que permiten establecer estrategias de difusión de la información en las organizaciones; y análisis de patrones -se interpretan y validan los patrones1-.
Se realiza un estudio webmétrico de uso a la revista electrónica Avanzada Científica a través de sus ficheros log entre los años 2011 y 2014. Concretamente se investiga desde el 22 de septiembre de 2013 al 23 de septiembre de 2014 y se compara con un anterior trabajo que se enmarcó entre el 5 de septiembre de 2010 al 28 de julio de 2011. Se realiza una amplia búsqueda de información en la web. Se aplica el software Sawmill 7. La comparativa se realiza a los indicadores de visibilidad: análisis de enlaces, accesos a páginas, por usuario, por país de origen, petición, visitas, tipo de navegador y sistema operativo utilizado para consultar la revista, cantidad de hits, tipos de errores y sus códigos de identificación, ciudades, días de la semana y horarios entre otros no menos importantes.	Minería web de contenido: La minería de contenido del web trata de extraer información relevante sobre el contenido de la web de manera que pueda ayudar clasificarlo, aumentando la organización de ese contenido, para posteriormente mejorar el acceso y la recuperación de la información en él contenida. La Minería web de estructura que se usa para conocer cómo está organizada una web, cómo está estructurada y cómo es la navegación a través de ella y la Minería de uso del web que trata de extraer patrones de uso del web por parte de los usuarios. Para ello se utilizan los archivos .log de los servidores Web de forma que aplicando minería de textos sobre ellos se pueda extraer información útil. Este tipo de minería tiene como objetivos principales: identificar patrones generales de uso de un sitio web de manera que se pueda reestructurar para que sea más fácil de utilizar y mejore el acceso por parte de los usuarios, y obtener perfiles de los distintos tipos de usuarios a través de su comportamiento y navegación, para poder atender de forma más personalizada.

CONCLUSIONES	PALABRAS CLAVES	Minería web	Contenido web	Minería web de contenido	Técnicas de minería web
Es importante entonces el 'buen diseño' en las interfaces digitales, lo que según Galitz (2007) puede llegar a aumentar la productividad de los usuarios de un 25% a un 40% e impactar en los aspectos económicos de las empresas. Entre muchos otros aspectos, menciona el autor, es necesario conocer las necesidades de los usuarios y entender los objetivos de las interfaces para que las tareas de los usuarios no resulten frustrantes y/o fatigosas. Los sitios web son productos que deben explicarse a sí mismos ante el usuario, independientemente del tipo de sitio que sea; deben poder facilitar la ejecución de tareas dentro de sí mismos y no volver frustrante el proceso (Garret, 2011). Por ello, es necesario investigar las necesidades de los usuarios y diseñar interfaces a su medida, que puedan ser útiles para los objetivos que se quieren crear (Chisnell & Redish, 2005).	Usabilidad; interacción humano-computador.		x		
Como puede verse, el trabajo para indexar en los sitios en la Deep Web resulta más complicado desde su consulta hasta su almacenamiento, pero también se han realizado trabajos que permiten encontrar estos tipos de páginas por la importancia y valor que representa la información que contienen. Pero aun así existe un vacío que consiste en las páginas contenidas en la Deep Web las cuales no es posible indexar, ya que se hace uso de una tecnología llamada TOR (The Onion Routing). Ésta es una red abierta que le permite a los usuarios defenderse contra el análisis de tráfico que realizan algunas instancias gubernamentales sobre la Internet, y es una forma de vigilancia que amenaza la libertad personal y la privacidad, confidencialidad en los negocios, como de las relaciones, y la seguridad del Estado	Web oculta, web de la superficie, rastreadores.	x	x	x	
En este editorial, presentamos una visión de la minería de datos en el comercio electrónico, y describimos algunos de los problemas que pueden ser resueltos por las técnicas detrás de este campo. También hemos proporcionado una visión general de algunos de los técnicas de aprendizaje automático más comúnmente utilizadas en la minería de datos. Como ejemplo, hemos elegido describir el La asociación rige la técnica de minería en más detalle y ha hablado sobre algunos de los problemas asociados con su solicitud. En el futuro, visualizamos proporcionar descripciones y discusión de otras técnicas en la minería de datos y cómo se pueden aplicar al comercio electrónico. En resumen, la minería de datos juega un papel importante en el desarrollo de aplicaciones de comercio electrónico. Electrónico el comercio y los campos relacionados con la inteligencia empresarial y el análisis se han desarrollado en gran medida debido a la madurez de minería de datos y otras áreas. Estas áreas de investigación se han vuelto más importantes debido a la llegada de grandes datos y la posterior iluminación de la larga cola en el comercio electrónico.		x	x		x
Aplicar técnicas de minería de datos requiere de una precisa transformación de los mismos; en este caso, esa fue la etapa más costosa en tiempo, conforme a la irregularidad en la calidad de dichos datos, ya que los registros presentaban incoherencias y una gran cantidad de los datos de cierto atributo no eran persistentes. Igualmente, se encontraron muchos datos nulos o faltantes y otros redundantes. Algunas variables categóricas, como el estrato del egresado, no representaron un papel importante en los algoritmos de clasificación. Adicionalmente, el tipo de contrato y la relación entre la carrera y el puesto de trabajo son influyentes en la percepción de la calificación de las habilidades y destrezas adquiridas en los estudios.	Egresados, Minería de datos, Técnicas de clasificación.	x			x
El proceso de obtención de conocimientos de las bases de datos se compone de varias fases: fase de integración y limpieza, fase de selección y tyransformación, fase de minería de datos y por último, fase de evaluación e interpretación. Esto se aplica sobre un volumen de datos muy amplio, buscando obtener la información de los datos en este proceso iterativo e interactivo					x
La metodología CRISP-DM demostró ser una metodología poderosa ya que esta abarca todo el ciclo de vida del proyecto de minería de datos, adicionalmente demostró una gran adaptabilidad a las necesidades de la CCAQ teniendo en cuenta que un proyecto de este tipo no tiene precedentes dentro de la entidad. Si bien CRISP-DM no es una fórmula mágica para el éxito de un proyecto de minería de datos, si se involucra algo de formación y algunos expertos, da un excelente punto de inicio como herramienta para dar respuesta a las preguntas que tengan los directivos sobre el negocio.		x			x
Si bien es cierto que esta imbricación entre los métodos computacionales y otras disciplinas supone cambios en el quehacer científico, los big data y las herramientas relacionadas invitan a repensar las lógicas de investigación social y del propio periodismo desde una perspectiva más amplia, donde se desdibujan aún más los límites entre los campos de estudio y de obtención de información. Las nuevas lógicas implican la necesidad de construir equipos interdisciplinares y centros de análisis de big data en las universidades y centros de investigación, que faciliten el desarrollo de proyectos de investigación para explotar el enorme potencial de análisis de estas fuentes para las ciencias sociales y el periodismo.	Datos; Big data; Minería de datos; Aprendizaje automático; Modelamiento de temas; Análisis de sentimientos	x			
En vista del avance tecnológico, el cual poco a poco acoge en su manto, diferentes campos de estudio, es lógico afirmar que con la ayuda del desarrollo tecnológico, las invenciones y descubrimientos se componen y producen progresivamente de manera exponencial. Cuando se desease relatar, expresar, explicar o mostrar una propuesta, esta debe estar soportada en archivos y trabajos previos verídicos, siendo así entonces, una forma de colaborar con el progreso del estudio, la implementación de la minería de texto a un punto analítico, donde con su metodología y técnica, se puedan tomar decisiones más acertadas sobre qué tipo de material será cimiento o soporte de un trabajo o propuesta venidera. Aplicando la minería de texto, se analizarán diferentes estructuras de datos que comúnmente se usan como sustento argumentativo en los trabajos o redacciones, de tal forma que se pueda conocer que tan acertado y útil es el soporte en análisis, siendo así, una referencia de carácter significativo, donde su contenido puede ser utilizado para generar un descubrimiento o nuevo conocimiento con más precisión.	Minería de texto, Minería de datos, Estructura de datos	x			x
Actualmente, estamos extendiendo medidas de calidad basadas en informaci´on factual, de manera tal de detectar fallas de calidad espec´ificas. En este contexto, se est´an realizando pruebas con el subconjunto de fallas de calidad de Wikipedia en ingl´es denominado Original Research (una de las diez fallas m´as importantes, mencionadas precedentemente) para determinar la efectividad de este tipo de features.	Miner´ıa de Textos, Miner ´ıa de la Web, Ling¨u´ıstica Computacional, Procesamiento del Lenguaje Natural		x	x	

La clasificación de los tweets según su zona geográfica ha estado fuertemente condicionada por la escasez de tweets que incluyan su geolocalización. La solución a este problema ha llevado a la obtención del lugar del usuario que aparece en el perfil, con la pérdida de exactitud que esto supone. No obstante, dado que muchos usuarios no indican un lugar válido tampoco en su perfil, el número de tweets geolocalizados en territorio nacional sigue siendo pequeño.	Twitter, minería de datos, recuperación de información, geolocalización, procesado de lenguaje natural		x	x	
Con el desarrollo, implementación y presentación de los resultados obtenidos de este trabajo se resalta la suma importancia del conjunto de técnicas computacionales propuesto para las formas de marketing y el desarrollo de organizaciones y empresas. Se realizo la extracción de comentarios en la red social y en Amazon utilizando el rastreo Web como medio computacional y posteriormente utilizar técnicas que permitan hacer un análisis connotativo de manera ordenada y metodológicamente siguiendo un proceso el cual arroja un resultado que puede ser usado e implementado por departamentos de marketing que ofrezcan o tengan vinculado su sector a exposición de productos a cualquier página de compras en línea que se encuentran en Internet.	Inteligencia artificial, Redes neuronales, Machine learning		x	x	
El desarrollo del proyecto permite definir a la extracción y almacenamiento de datos de redes sociales como una práctica o conjunto de procesos que implica la comprensión de una base teórica referente a arquitectura de software, programación y almacenamiento de datos, ya que sin esta base conceptual el entorno de desarrollo de una herramienta que permita establecer estas funcionalidades es invisible a los ojos del desarrollador. Los tipos de datos recolectados en una extracción a un medio definido pueden variar, y esta estructura variable entre ellos implica el uso de diferentes técnicas de manejo de datos y sistemas de almacenamiento. El correcto tratamiento de estos datos puede definir en su totalidad los resultados obtenidos al culminar el proceso de desarrollo. En este caso modelos relacionales y no relacionales fueron integrados en el mismo sistema, con el fin de alcanzar el nivel de calidad y cumplir con los objetivos del proyecto, encontrando diferentes ventajas al definir correctamente la estructura general del desarrollo.	Almacenamiento de datos, BigData, Desarrollo de software, Extracción, Minería de datos	x			
En este trabajo se propone el uso de agrupamiento o clustering para mejorar la minería de procesos educativa y, al mismo tiempo, optimizar tanto el rendimiento/ajuste y comprensibilidad/tamaño del modelo obtenido. La comprensibilidad del modelo obtenido es un objetivo básico en la educación, debido a la transferencia de conocimientos básicos que ello conlleva. Realizar gráficos, modelos o una representación visual más accesible o al menos, accesible, para los profesores y estudiantes, hacen que estos resultados sean muy útiles para el seguimiento del proceso de aprendizaje y para proporcionar una retroalimentación, siendo uno de nuestros futuros retos realizarlo en tiempo real. Además, Moodle no proporciona herramientas de visualización específicas de los datos usados por los estudiantes que permitan a los diferentes agentes del proceso de aprendizaje entender estas grandes cantidades de datos "en bruto" y, tomen consciencia de lo que esta pasando en una educación a distancia, además de ampliar el uso de los resultados de Entornos de Aprendizaje Hipermedia Adaptativos en los que es muy útil motivar a los estudiantes o recomendarles rutas de aprendizaje, con el fin de mejorar la experiencia de aprendizaje de una manera más estratégica.	Base de datos; aprendizaje; estudiante; red de información	x	x	x	x
Las aplicaciones gráficas se utilizan a menudo para mostrar más claramente resultados de la minería de textos. Han ido avanzando a lo largo de estos últimos años, llegando a dejar al usuario ejecutar un cierto número de consultas preprogramadas y fijas. Actualmente, se ha conseguido que estos sistemas de minería de textos puedan exponer gran parte de su funcionalidad al usuario, mediante un acceso a las líneas de comando del programa. Con la finalidad de que el trabajo realizado se pudiese ver e interactuar con él con la máxima sencillez posible se creó una aplicación Shiny del clasificador de textos utilizando RStudio.	Minería web; minería de texto			x	x
El uso de técnicas y herramientas de Data Mining aplicadas a bases de datos musicales permite encontrar ciertas asociaciones entre las preferencias musicales de un grupo de personas y rasgos en su personalidad, aunque estas asociaciones no son concluyentes. La metodología SEMMA de Data Mining a bases de datos de contenido musical ofrece una secuencia sistemática de pasos adecuados para obtener resultados objetivos en tareas de minería de datos.	Música, personalidad, mineríad e datos, modelo Ocean, inteligencia de negocios	x	x	x	x
En términos globales se observó un valor proporcionalmente bajo de titulares que hablaban directamente sobre México: 3842 noticias que equivalen al 0,6% del total de noticias emitidas por los 290 medios seguidos. Se trata de una cantidad promedio cercana a las menciones que se hacen de otros países iberoamericanos	Tratamiento Informativo, Noticias, Minería de datos, México, Chile	x	x	x	x
La utilización de técnicas de Minería de Datos para el análisis de la lealtad de los clientes dentro de la empresa tecnológica Master PC, le permitirá elaborar estrategias de retención hacia sus clientes, en lugar de pagar un alto costo para la atracción de nuevos clientes.	Minería de datos, segmentación, empresa tecnológica	x			x
Dentro del dataset se tomó los campos de Body y Score, siendo body el campo de comentarios y Score el puntaje o Karma en Reddit. Del mismo modo, se utilizó las variables temporales conteo y media. Donde, conteo es el identificador de los comentarios y bueno, media es la media total de los comentarios. El uso de algoritmos de Machine Learning como lo son las Redes Neuronales Artificiales para el análisis de información dentro de grandes bases de datos siempre será determinado bajo el enfoque que conlleve el análisis o estudio, en este caso se lo efectuó con una modalidad investigativa para demostrar que puede resolver cualquier tipo de problemática y a la vez facilitar la resolución de problemas a futuro.	DataMining, Artificial Neural Network, Python, Artificial Intelligence, DeepLearning	x	x		
Esta investigación supone un instrumento de trabajo para el webmaster del site estudiado, con el cual puede mejorar el posicionamiento SEO. Con este análisis queda demostrado que es necesario utilizar los descriptores adecuados para aumentar el posicionamiento del sitio web. Para ello es recomendable incluir, de manera estratégica, los diez términos analizados, en las distintas zonas del site (título, encabezados, URL y otros).	Minería web, sitio web, metadatos, AWStats, motores de búsqueda, términos de búsquedas.	x	x	x	
En el ámbito de la ciencia y la innovación tecnológica, la medición de sus resultados por los llamados indicadores de impacto □cuyo principal objetivo es evaluar el beneficio tangible, la repercusión del resultado y no el resultado en sí□ constituye un estadio superior en la evaluación de la producción científica e innovadora de investigadores y tecnólogos cubanos. Por lo que este estudio webmétrico de uso refuerza estos indicadores de impacto como el de autores y su calidad profesional, cantidad de visitante, cantidad de países que nos visitan, a la cantidad de empresas que publican las cuales se relacionan con las empresas priorizadas en el territorio, las temáticas tratadas, también se relacionan con los proyectos de desarrollo científicos productivos, aunque no sean objetivo de este estudio, pero que están contenido.					